Извлечение информации (information extraction, IE)

Павел Браславский

Задача

- Неструктурированный текст → структурированное представление части информации
- Источники:
  - Новости
  - Википедия
  - Резюме
  - Научные статьи (биология, медицина)
  - Художественные произведения
- Результат
  - виден пользователям
  - для использования в других приложениях (поиск, вопросно-ответный поиск, …)

Приложения

- аналитика
- интерфейс
- анализ тональности
- поиск
  - вопросно-ответный поиск (QA)
  - поиск сущностей (entity search)

ПРИМЕРЫ

годовая сумма выплат за 2013-2015 гг. составляла как минимум 75% свободного денежного потока оператора, но не менее 40 млрд руб.

МТС превзошла конкурентов по росту выручки от мобильной связи

Снижение капитальных затрат может хорошо отразиться на ее дивидендах

Таким образом, МТС отказывается от привязки дивидендов к свободному денежному потоку в пользу конкретного показателя выплаты на акцию. Этот переход должен обеспечить стабильный доход акционеров, пояснил сегодня журналистам вице-президент МТС по финансам и инвестициям Алексей Корня. По его словам, 25-26 руб. на акцию - это исторически максимальный уровень дивидендной доходности оператора. Выплаты на этом уровне компания хочет сохранить в течение следующих трех лет. При текущем количестве акций МТС это предполагает общую сумму дивидендных выплат в 50-52 млрд руб., добавляет Корня. Рекордными для МТС стали дивиденды, выплаченные по итогам 2014 г., - тогда ее акционеры получили 53,3 млрд руб., что составило 93,5% свободного денежного потока оператора. О том, что МТС в 2016 г. сохранит дивидендные выплаты на уровне прошлого года, 7 апреля заявил президент «АФК Система» (основной акционер МТС) Михаил Шамолин.

Кроме того, руководство МТС оставляет за собой возможность выкупить и погасить акции на 30 млрд руб. в течение трех ближайших лет, говорится в сообщении. У МТС нет обязательства совершать эту сделку, речь идет о возможности рассмотрения такого решения руководством компании, отмечает Алексей Корня. По его словам, МТС не относит выкуп акций к одному из условий выполнения прогноза по дивидендам, а рассматривает лишь как дополнительную опцию, которая позволит в случае благоприятных финансовых условий «заплатить акционерам еще больше». В этом случае сумма выплачиваемых дивидендов в абсолютном значении уменьшится, но доходность на акцию сохранится на прогнозируемом уровне - в 25-26 руб, отметил Корня.

Совет директоров рекомендовал собранию акционеров утвердить дивиденды за 2015 г. в размере 14,01 руб. на акцию (28,02 руб. на одну ADR). Общая сумма выплат акционерам в этом случае может составить 28 млрд руб. Выплатить эти деньги акционерам МТС предполагает до 1 августа 2016 г. МТС переходит к более равномерному распределению дивидендных выплат в течение года, говорится в сообщении компании. Год назад промежуточные дивиденды МТС за первое

Яндекс Новости

Главные новости | Мои новости | Политика | Общество | Экономика | В мире | Спорт

Адвокат опроверг слухи об УДО бойца ММА Емельяненко

С чего всё началось

Александр Емельяненко вышел из колонии по УДО

Боец смешанных единоборств Александр Емельяненко, осужденный за насильственные действия сексуального характера, освобожден из колонии условно-досрочно. Об этом сообщил ТАСС источник, знакомый с ситуацией. ТАСС 13:18

Боец ММА Александр Емельяненко продолжает отбывать наказание в колонии, сообщил агентству "Р-Спорт" адвокат спортсмена Кахабер Долбадзе. Ранее источник РИА Новости в правоохранительных органах сообщил, что Емельяненко освободился из колонии условно-досрочно. РИА Новости 14:16

ПОДРОБНЕЕ О СОБЫТИИ

РОССИЙСКИЕ ДРУГИЕ ССЫЛКИ

Вечерняя Москва
14:14

Боец Александр Емельяненко досрочно освобожден из колонии

Яндекс Новости

Главные новости | Мои новости | Политика | Общество | Экономика | В мире | Спорт

Адво
ММА

Александр Емельяненко

Родился: 2 августа 1981 г. (35 лет), Старый Оскол, Белгородская область, СССР

Российский боец смешанных единоборств. Бывший чемпион мира по версии ProFC. Многократный чемпион России и мира по боевому самбо, чемпион Европы по боевому самбо, мастер спорта России по самбо, мастер спорта России международного класса по боевому самбо, мастер спорта России по дзюдо. Википедия

Найти больше в Яндексе

С чего

Алек

Боец см
действи
сообщи

Боец ММА Александр Емельяненко продолжает отбывать наказание в колонии, сообщил агентству "Р-Спорт" адвокат спортсмена Кахабер Долбадзе. Ранее источник РИА Новости в правоохранительных органах сообщил, что Емельяненко освободился из колонии условно-досрочно. РИА Новости 14:16

Яндекс Новости

Главные новости | Мои новости | Политика | Общество | Экономика | В мире | Спорт | Происшествия | Культура | Наука | Те

Адвокат опроверг слухи об УДО бойца ММА Емельяненко

С чего всё началось

Александр Емельяненко вышел из колонии п

Боец смешанных единоборств Александр Емельяненко, осуж
действия сексуального характера, освобожден из колонии усл
сообщил ТАСС источник, знакомый с ситуацией. ТАСС 13:18

Боец ММА Александр Емельяненко продолжает отбывать н
сообщил агентству "Р-Спорт" адвокат спортсмена Кахабер
РИА Новости в правоохранительных органах сообщил, что Емельяненко освободился из колонии условно-досрочно. РИА Новости 14:16

Федор Емельяненко

Родился: 28 сентября 1976 г. (40 лет), Рубежное, Луганская область, Украинская ССР, СССР

Российский спортсмен, четырёхкратный чемпион мира по ММА в тяжёлом весе по версии «Pride FC», двукратный - по версии «RINGS», двукратный - по версии «WAMMA», четырёхкратный чемпион мира и девятикратный чемпион России по боевому самбо. Заслуженный мастер спорта по самбо и мастер спорта международного класса по дзюдо. Википедия

Найти больше в Яндексе

ОБЩЕСТВО

в новостях

Гайдар Мария Егоровна

дочь политика (219 упоминаний в СМИ)
...Как сообщила РИА "Новости" *дочь политика Мария Гайдар*, ее отец лежит под капельницей по восемь часов в день.... 01.12.06 СМИ.ru

молодежный политик (5 упоминаний)
... *Молодежные политики Маша Гайдар* и Илья Яшин висят под мостом... 23.11.06 Клерк.Ру

Работа

Места работы отсортированы по количеству упоминаний

Демократическая Альтернатива, лидер (347 упоминаний в СМИ с 10.06.2005 по 08.12.2006)
...23 ноября *лидер* молодежного движения *"Демократическая Альтернатива" Мария Гайдар* и Илья Яшин с помощью альпинистского снаряжения спустились с Большого Каменного моста и больше часа продержали десятиметровый транспарант с лозунгом "Верните народу выборы, гады" над Москвой-рекой.... 24.11.06 Томский Обзор

СПС, член (57 упоминаний с 12.06.2005 по 23.11.2006)
...Ежедневный журнал, 23.11.06Никита Белых, лидер партии СПС (Союз Правых Сил), об акции протеста против выборного законодательства, одним из организаторов которой стала *Мария Гайдар*, лидер молодежного движения "Да!" и *член* партии *СПС*: Тут два момента по существу вопроса- очевидно, что мы полностью поддерживаем

Связанные люди

Егор Гайдар
...- Мария Гайдар, дочь бывшего премьер-министра РФ Егора Гайдара.... Эхо Москвы в Перми

Подписка на новости

Новости на вашем сайте

пресс-портреты Яндекс.Новостей (2006-2012)

Случаются ошибки :)

Яндекс меня похоронил :)

Гайдар Мария Егоровна

Дата смерти — 23.11.2006
... Лидер молодежного "Яблока" Илья Яшин и координатор молодежного движения "Да" Мария Гайдар повесились сегодня в полдень под Большим Каменным мостом.... 23.11.06

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

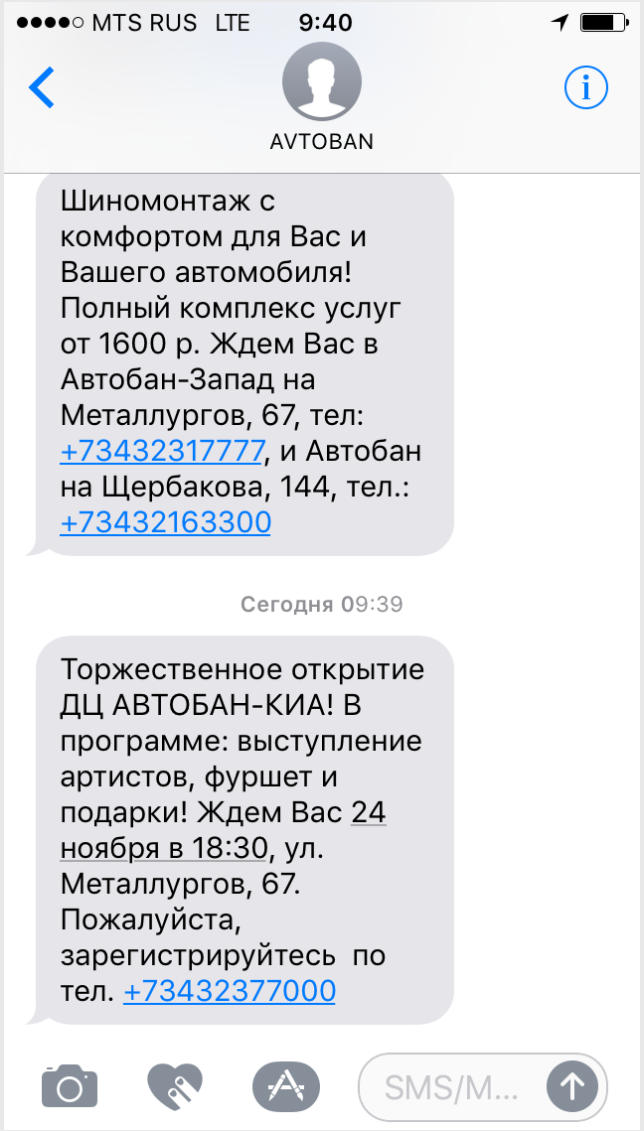
m_gaidar
8 декабря, 2006

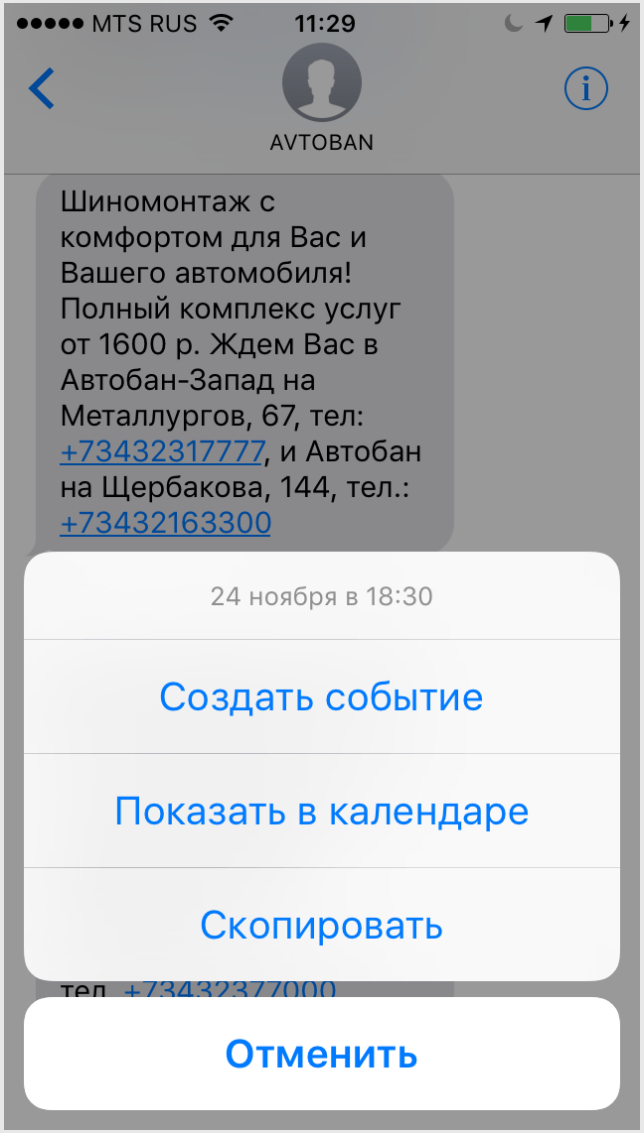
24 часа		
	1. Tele2	15
↑ !	2. оператор Tele2	5
↑ !	3. mvno	5
↑ !	4. Мегафон	4
↑ !	5. Ростелеком	4
↑ !	6. роскомнадзор	4
↑ !	7. Группа ВТБ	4
↑ !	8. Альфа-групп	3
↑ !	9. компания tele2	3
↑ !	10. ВТБ	3
↑ !	11. Альфа	3
↑ !	12. МТС	3
↑ !	13. Гамма	3
↑ !	14. пресс-служба " Ростелеком	3
↑ !	15. компания - Иркутский научно-исследовательский институт	3
↑ !	16. Turkcell	3
↑ !	17. АО " Иргиредмет	3
↑ !	18. филиал " Урал	3
↑ !	19. СТН	3
↑ !	20. компания Turkcell	3
↑ !	21. Henri Service Ltd	2
↑ !	22. Nadash International Holdings	2
↑ !	23. альянс " МегаФон	2
↑ !	24. пресс-служба АО " Иргиредмет	2

Неделя (02.11.16 - 08.11.16)		
	1. Tele2	468
↑ +10	2. тв	143
↑ +6	3. delfi	138
↑ +4	4. фото	137
↓ -1	5. delficообщите	137
↑ !	6. концерн Eesti Energia	134
↑ +11	7. Коммерческий	134
↑ +2	8. РЖД	134
↑ !	9. Eesti Raudtee	134
↑ !	10. h&m	134
↑ !	11. The Global Service Design Award	134
↑ !	12. taxify	134
↑ +4	13. центр Arsenal	134
↓ -9	14. ECA	134
↑ !	15. keskus	134
↑ !	16. Айдар	133
↑ !	17. CETA	130
↑ !	18. Tesla	128
↑ !	19. Талант	126
↑ +1	20. Eesti Energia	120
↑ !	21. правление Atrium	120
↑ !	22. компания Linda Line	119
↓ -10	23. Samsung	118
↑ !	24. Банк Эстония	116

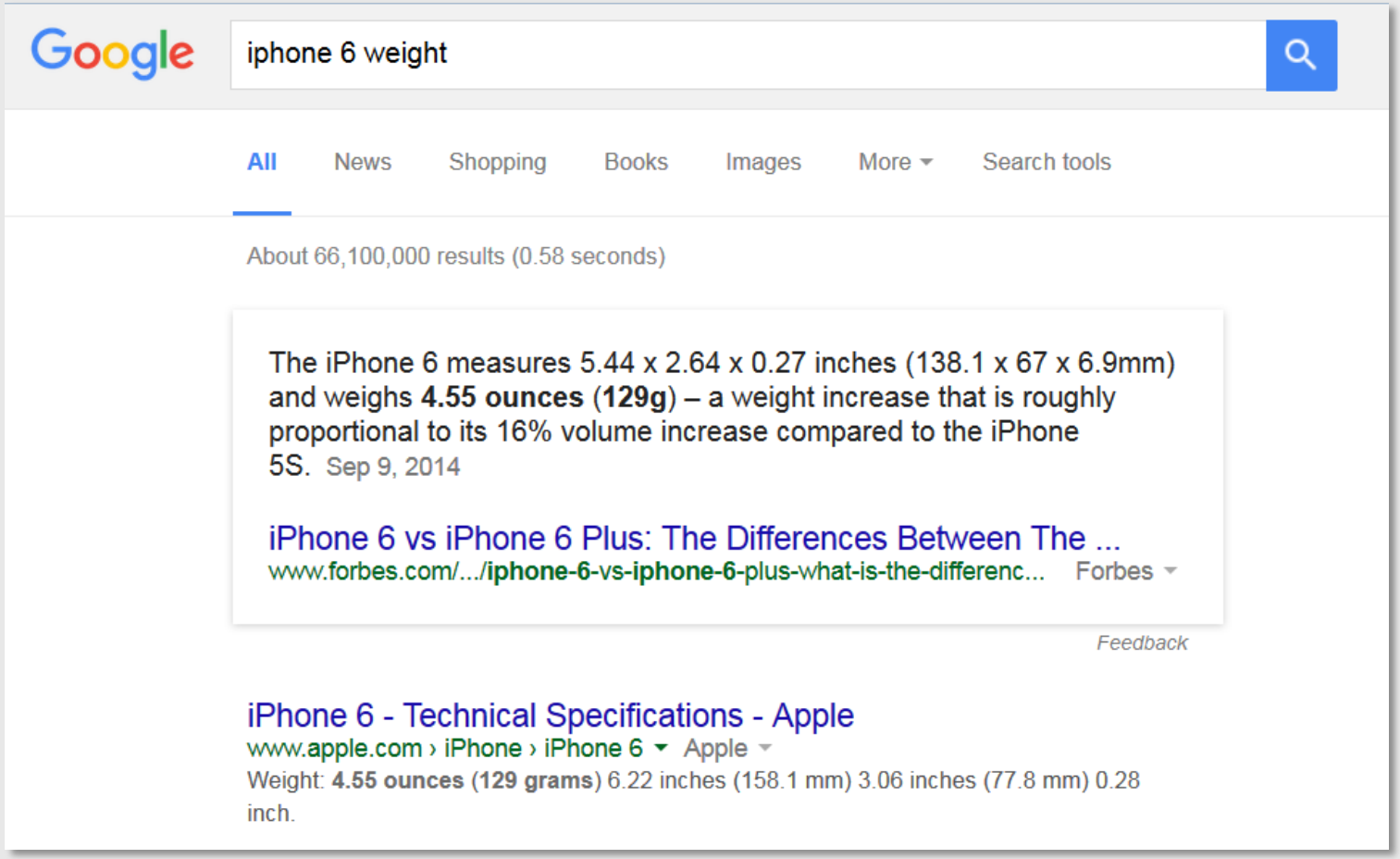
Месяц (09.10.16 - 08.11.16)	
1. Tele2	2576
2. Мегафон	532
3. МТС	498
4. delficообщите	357
5. ECA	354
6. сеть Tele2	345
7. филиал Tele2	339
8. фото	306
9. delfi	303
10. РЖД	281
11. Вымпелком	265
12. тв	265
13. Samsung	255
14. Билайн	249
15. Swedbank	246
16. Tallinn	234
17. центр Arsenal	229
18. Коммерческий	228
19. Marat	216
20. Какой	213
21. Eesti Energia	213
22. оператор Tele2	212
23. Ведомость	211
24. компания Tele2	208

https://br-analytics.ru/12371749/ner/#org


●●●●○ MTS RUS LTE 9:40
AVTOBAN
Шиномонтаж с комфортом для Вас и Вашего автомобиля! Полный комплекс услуг от 1600 р. Ждем Вас в Автобан-Запад на Металлургов, 67, тел: +73432317777, и Автобан на Щербакова, 144, тел.: +73432163300
Сегодня 09:39
Торжественное открытие ДЦ АВТОБАН-КИА! В программе: выступление артистов, фуршет и подарки! Ждем Вас 24 ноября в 18:30, ул. Металлургов, 67. Пожалуйста, зарегистрируйтесь по тел. +73432377000
SMS/M...

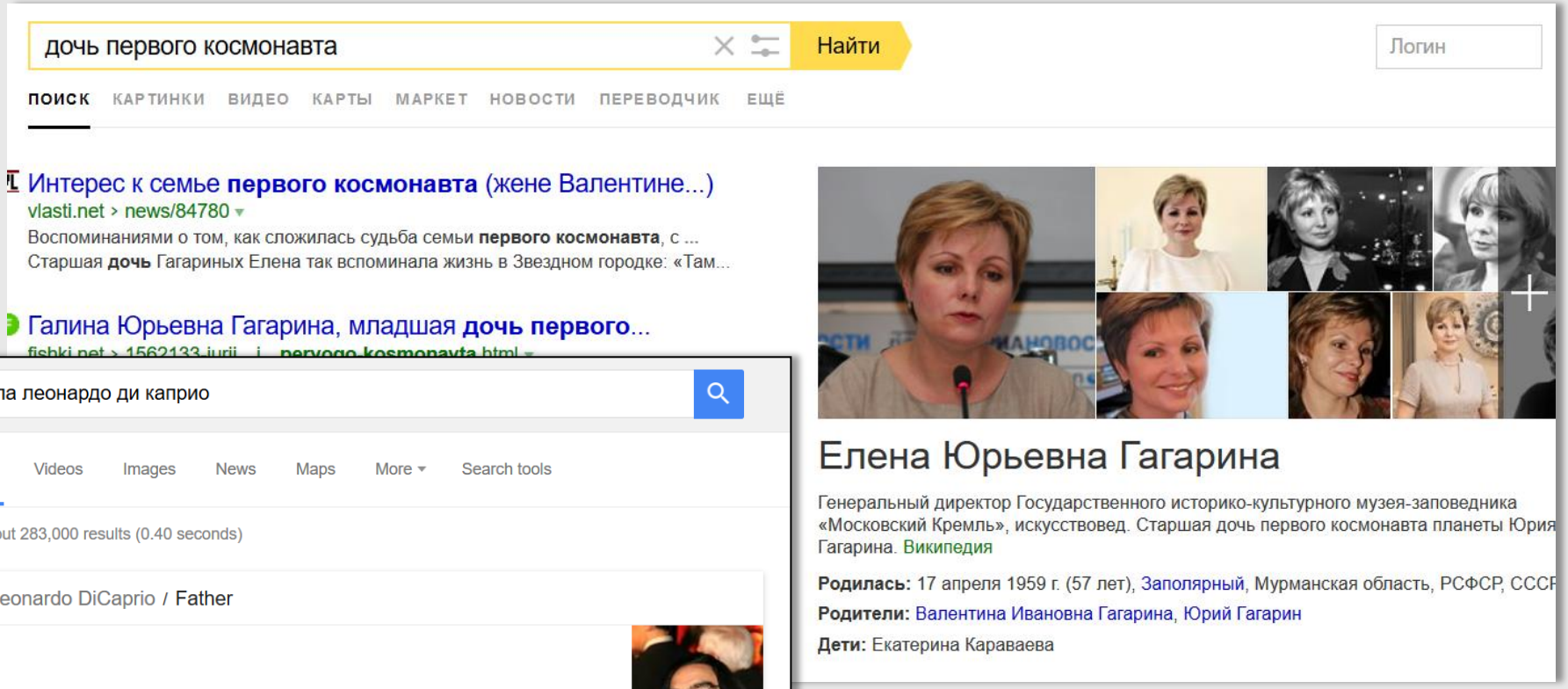

●●●●● MTS RUS 11:29
AVTOBAN
Шиномонтаж с комфортом для Вас и Вашего автомобиля! Полный комплекс услуг от 1600 р. Ждем Вас в Автобан-Запад на Металлургов, 67, тел: +73432317777, и Автобан на Щербакова, 144, тел.: +73432163300
24 ноября в 18:30
Создать событие
Показать в календаре
Скопировать
Отменить

Ответы



[из лекции об информационном поиске]

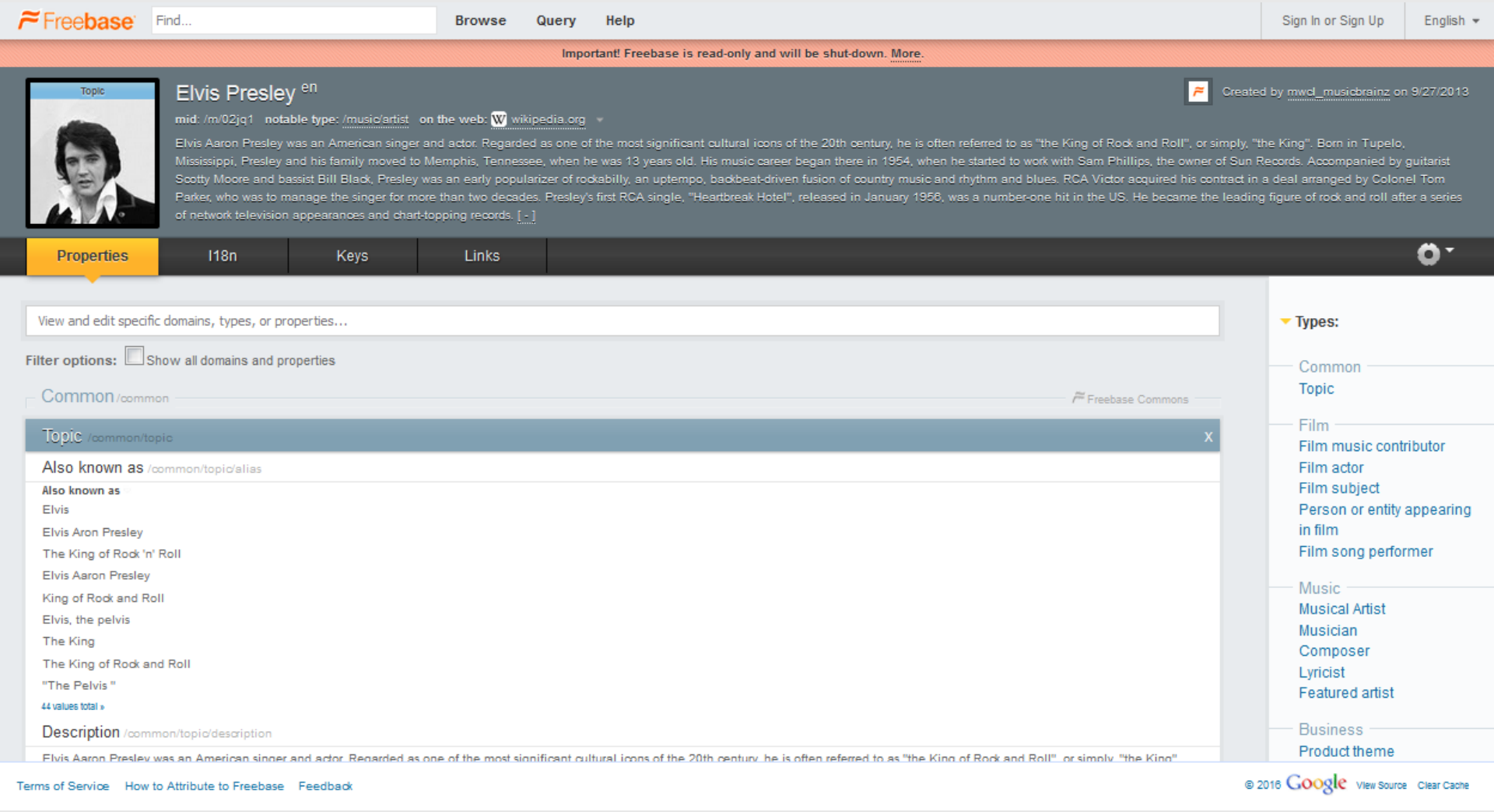
Поиск сущностей



[из лекции об информационном поиске]

БАЗА ЗНАНИЙ (KNOWLEDGE BASE)

Freebase


Freebase
Find...
Browse
Query
Help
Sign In or Sign Up
English
Important! Freebase is read-only and will be shut-down. More.
Topic
Elvis Presley en
Created by mwcl_musicbrainz on 9/27/2013
mid: /m/02jq1 notable type: /music/artist on the web: wikipedia.org
Elvis Aaron Presley was an American singer and actor. Regarded as one of the most significant cultural icons of the 20th century, he is often referred to as "the King of Rock and Roll", or simply, "the King". Born in Tupelo, Mississippi, Presley and his family moved to Memphis, Tennessee, when he was 13 years old. His music career began there in 1954, when he started to work with Sam Phillips, the owner of Sun Records. Accompanied by guitarist Scotty Moore and bassist Bill Black, Presley was an early popularizer of rockabilly, an uptempo, backbeat-driven fusion of country music and rhythm and blues. RCA Victor acquired his contract in a deal arranged by Colonel Tom Parker, who was to manage the singer for more than two decades. Presley's first RCA single, "Heartbreak Hotel", released in January 1956, was a number-one hit in the US. He became the leading figure of rock and roll after a series of network television appearances and chart-topping records. [-]
Properties
I18n
Keys
Links
View and edit specific domains, types, or properties...
Filter options:
Show all domains and properties
Common /common
Freebase Commons
Topic /common/topic
Also known as /common/topic/alias
Also known as
Elvis
Elvis Aron Presley
The King of Rock 'n' Roll
Elvis Aaron Presley
King of Rock and Roll
Elvis, the pelvis
The King
The King of Rock and Roll
"The Pelvis "
44 values total »
Description /common/topic/description
Types:
Common
Topic
Film
Film music contributor
Film actor
Film subject
Person or entity appearing in film
Film song performer
Music
Musical Artist
Musician
Composer
Lyricist
Featured artist
Business
Product theme
Terms of Service
How to Attribute to Freebase
Feedback
© 2016 Google
View Source
Clear Cache

About: Юрий Алексеевич Гагарин

Goto Sponge NotDistinct Permalink

An Entity of Type : http://www.wikidata.org/ontology#Item, within Data Space : dbpedia.org associated with source document(s)

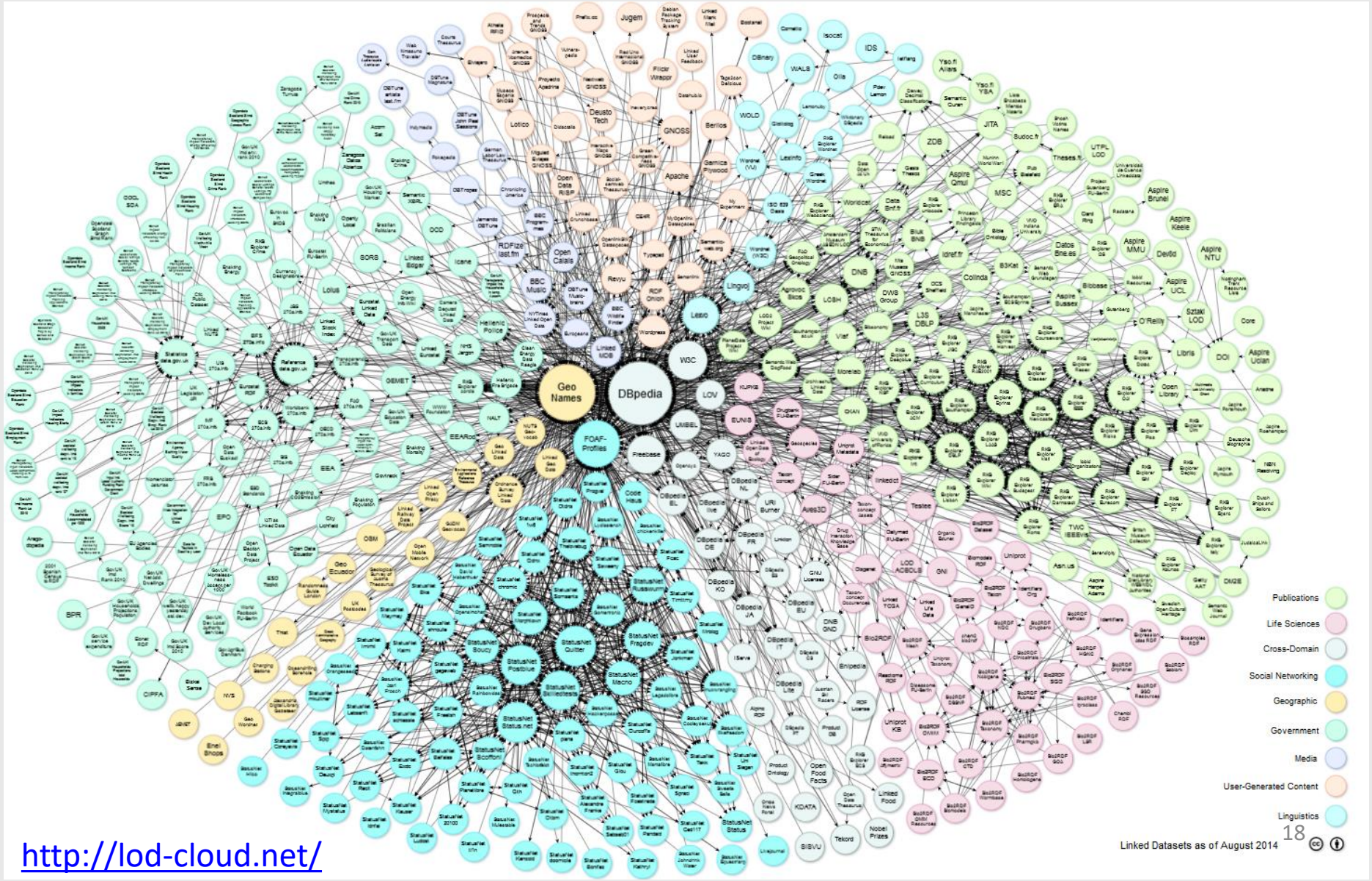
Type: http://www.wikidata.org/ontology#Item Command: Start New Facet Go

Attributes	Values
type	Человек http://www.wikidata.org/ontology#Item
wikidata:P1741c	100616
label	Yuri Gagarin Yuri Gagarin Yuri Gagarin

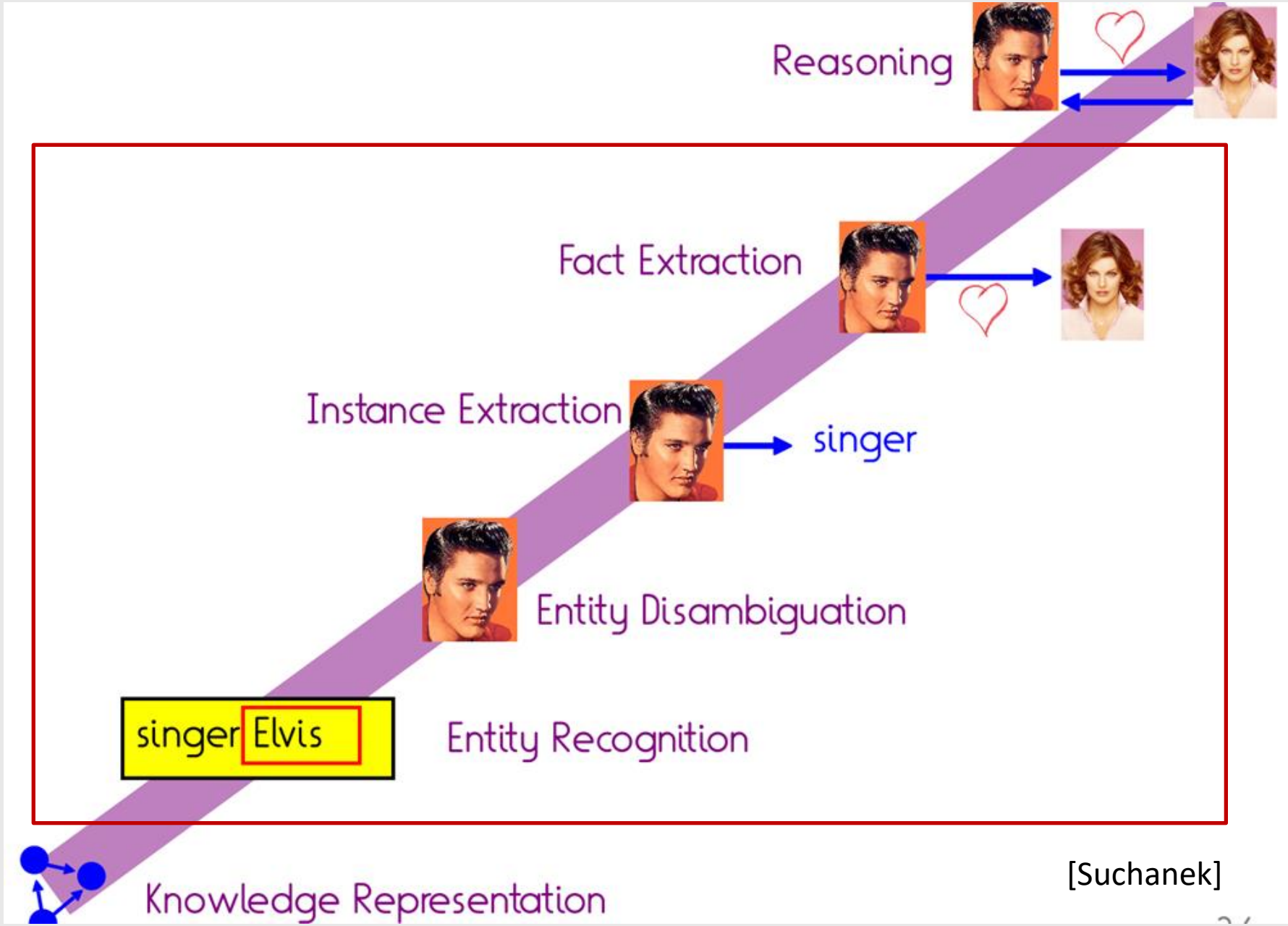
◂◂ First ◂ Prev Next ▸ Last ▸▸

wikidata:P119c	Некрополь у Кремлёвской стены
wikidata:P227c	118537105
wikidata:P241c	Центр подготовки космонавтов имени Ю. А. Гагарина Военно-воздушные силы СССР
wikidata:P245c	500342574
wikidata:P26c	Гагарина, Валентина Ивановна
wikidata:P40c	Гагарина, Елена Юрьевна
wikidata:P935c	Юрий Алексеевич Гагарин
wikidata:P535s	wikidata:Q7327S77B6D582-2641-4454-AA9A-A11CFE480837
wikidata:P1263c	666/000026588

Linked Open Data (LOD)



http://lod-cloud.net/


Reasoning
Fact Extraction
Instance Extraction
singer
Entity Disambiguation
singer Elvis
Entity Recognition
Knowledge Representation

[Suchanek]

ИМЕНОВАННЫЕ СУЩНОСТИ (NAMED ENTITIES)

Именованные сущности (NEs)

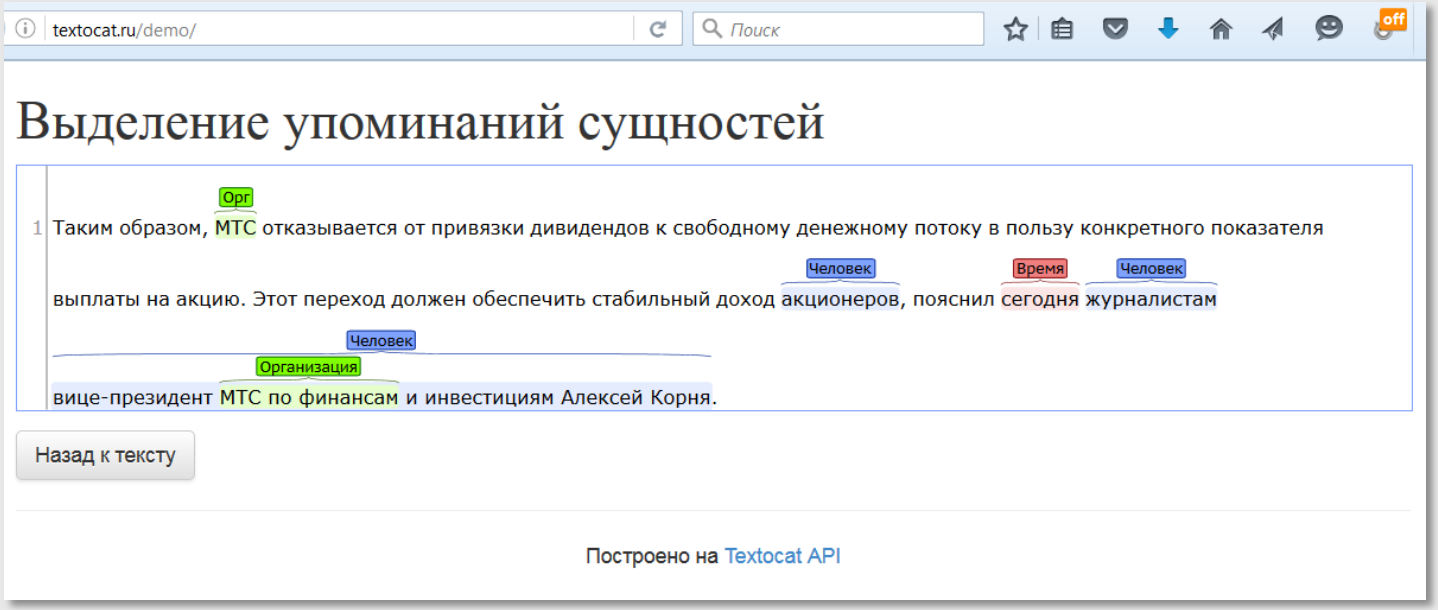
- Люди
- Организации
  - Фирмы
  - Правительственные организации
  - Газеты
  - Спортивные клубы
- Бренды
- Место (адрес, географический объект)
- События
- Даты
- Вес/цена/…

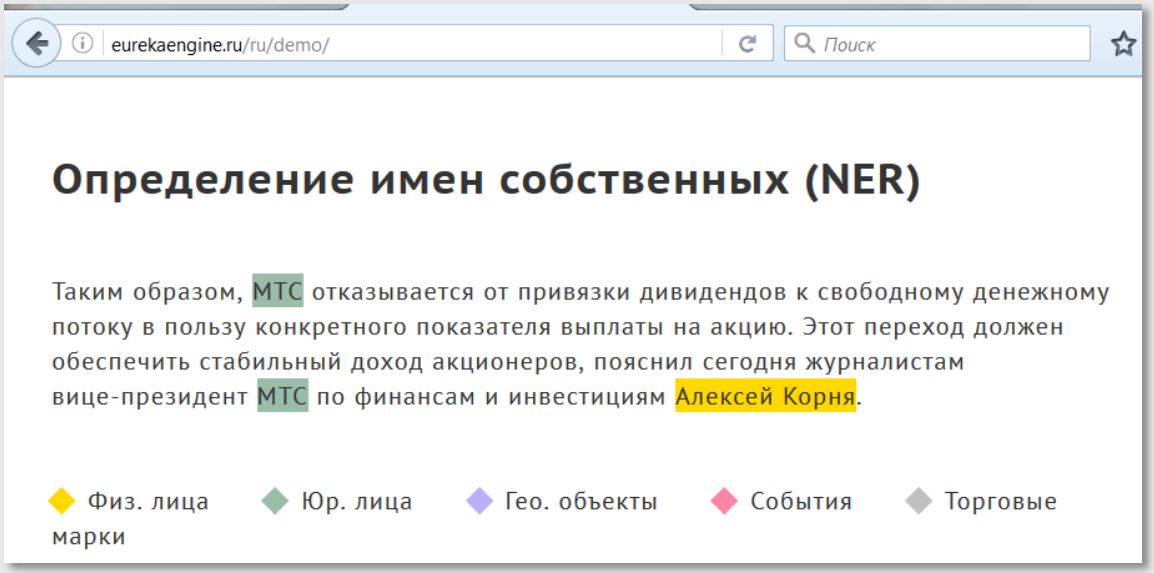
IO/BIO-разметка

Tag	Meaning
O	Not part of a named entity
B-PER	First word of a person name
I-PER	Continuation of a person name
B-LOC	First word of a location name
I-LOC	Continuation of a location name
B-ORG	First word of an organization name
I-ORG	Continuation of an organization name
B-MISC	First word of another kind of named entity
I-MISC	Continuation of another kind of named entity

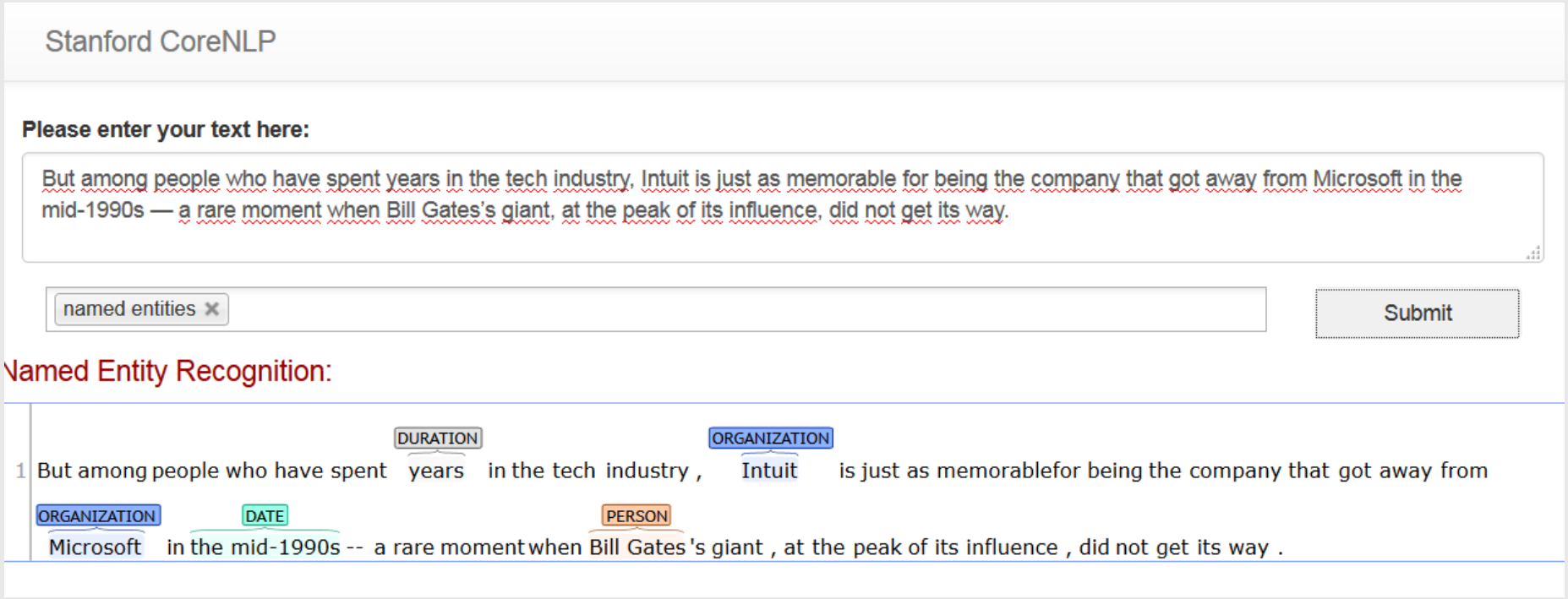
[Goldberg]

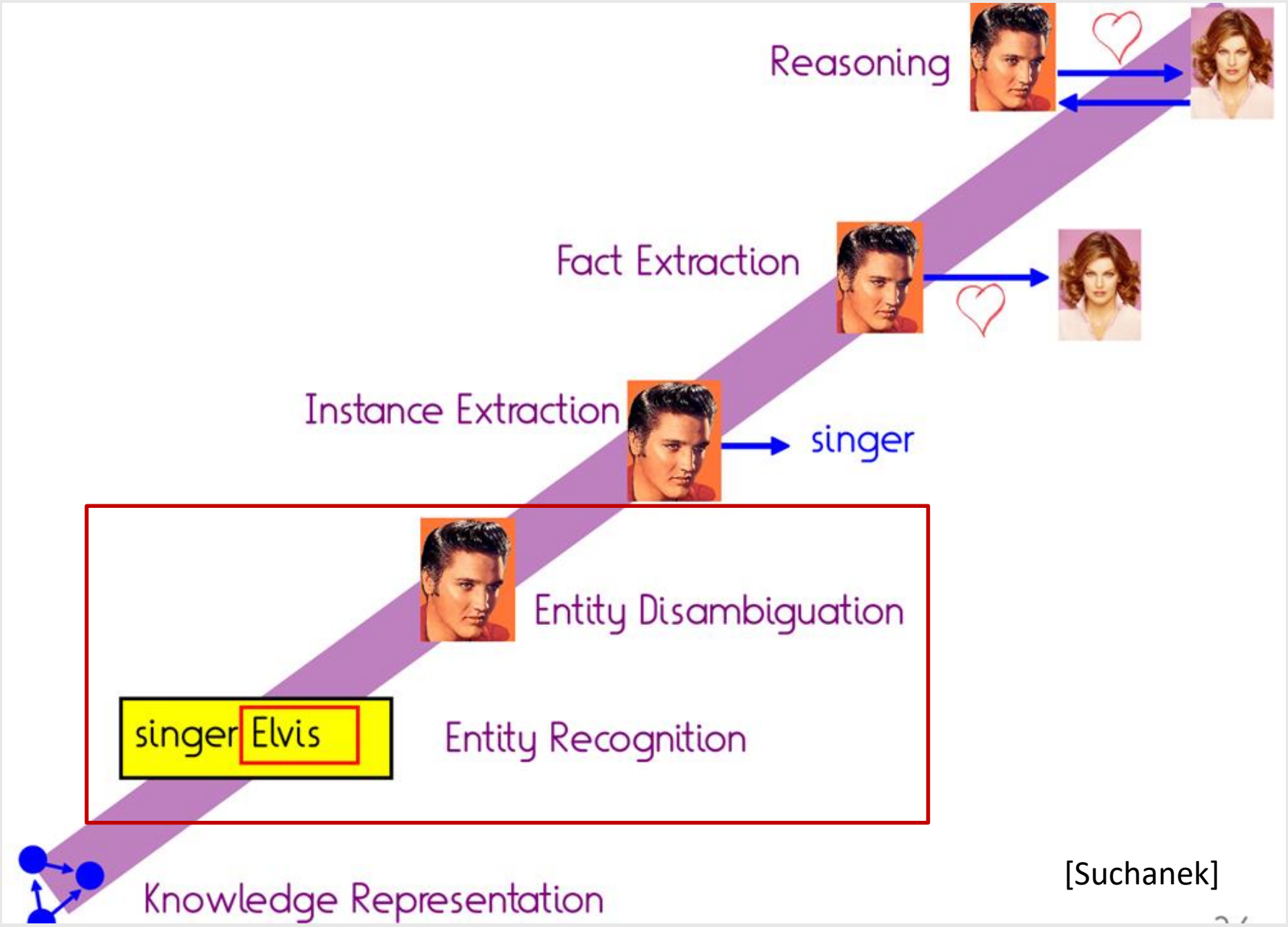
Пример


textocat.ru/demo/
Поиск
Выделение упоминаний сущностей
Орг
1 Таким образом, МТС отказывается от привязки дивидендов к свободному денежному потоку в пользу конкретного показателя
Человек
Время
Человек
выплаты на акцию. Этот переход должен обеспечить стабильный доход акционеров, пояснил сегодня журналистам
Человек
Организация
вице-президент МТС по финансам и инвестициям Алексей Корня.
Назад к тексту
Построено на Textocat API


eurekaengine.ru/ru/demo/
Поиск
Определение имен собственных (NER)
Таким образом, МТС отказывается от привязки дивидендов к свободному денежному потоку в пользу конкретного показателя выплаты на акцию. Этот переход должен обеспечить стабильный доход акционеров, пояснил сегодня журналистам вице-президент МТС по финансам и инвестициям Алексей Корня.
Физ. лица
Юр. лица
Гео. объекты
События
Торговые марки

Пример


Stanford CoreNLP
Please enter your text here:
But among people who have spent years in the tech industry, Intuit is just as memorable for being the company that got away from Microsoft in the mid-1990s — a rare moment when Bill Gates's giant, at the peak of its influence, did not get its way.
named entities
Submit
Named Entity Recognition:
1 But among people who have spent DURATION years in the tech industry , ORGANIZATION Intuit is just as memorablefor being the company that got away from
ORGANIZATION Microsoft in the DATE mid-1990s -- a rare moment when PERSON Bill Gates 's giant , at the peak of its influence , did not get its way .

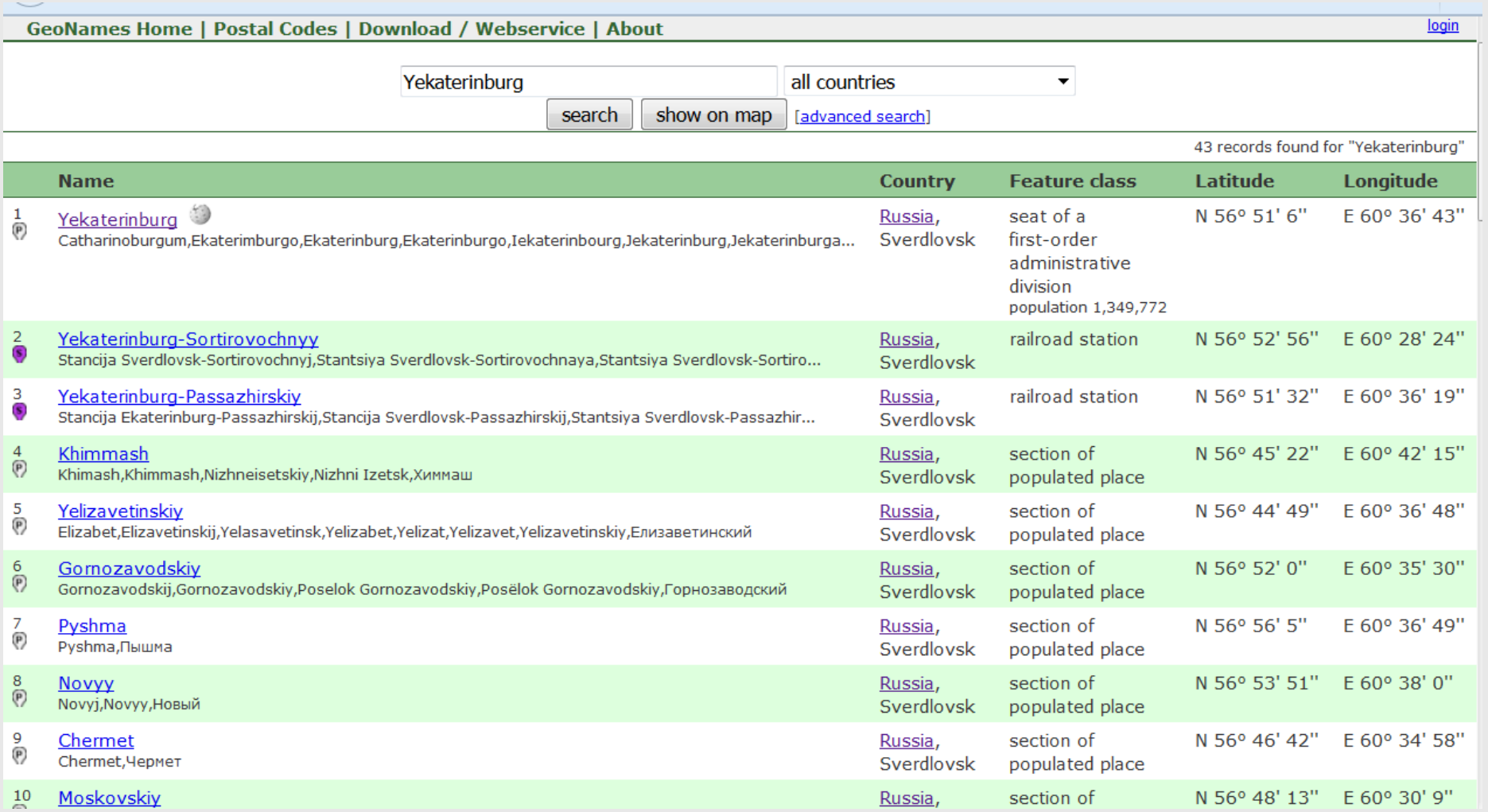

Reasoning
Fact Extraction
Instance Extraction
singer
Entity Disambiguation
singer Elvis
Entity Recognition
Knowledge Representation

[Suchanek]

Методы

- Регулярные выражения/шаблоны
- Словари (газетир, gazetteer)
- Классификация
- Методы на последовательностях:
  - Марковские модели
  - Условные случайные поля (Conditional Random Fields, CRF)
  - нейронные сети
- Обучение шаблонов

GeoNames



GeoNames Home | Postal Codes | Download / Webservice | About

login

Yekaterinburg | all countries

search | show on map | [advanced search]

43 records found for "Yekaterinburg"

	Name	Country	Feature class	Latitude	Longitude
1	Yekaterinburg Catharinoburgum,Ekaterimburgo,Ekaterinburg,Ekaterinburgo,Iekaterinbourg,Jekaterinburg,Jekaterinburga...	Russia, Sverdlovsk	seat of a first-order administrative division population 1,349,772	N 56° 51' 6"	E 60° 36' 43"
2	Yekaterinburg-Sortirovochnyy Stancija Sverdlovsk-Sortirovochnyj,Stantsiya Sverdlovsk-Sortirovochnaya,Stantsiya Sverdlovsk-Sortiro...	Russia, Sverdlovsk	railroad station	N 56° 52' 56"	E 60° 28' 24"
3	Yekaterinburg-Passazhirskiy Stancija Ekaterinburg-Passazhirskij,Stancija Sverdlovsk-Passazhirskij,Stantsiya Sverdlovsk-Passazhir...	Russia, Sverdlovsk	railroad station	N 56° 51' 32"	E 60° 36' 19"
4	Khimmash Khimash,Khimmash,Nizhneisetskiy,Nizhni Izetsk,Химмаш	Russia, Sverdlovsk	section of populated place	N 56° 45' 22"	E 60° 42' 15"
5	Yelizavetinskiy Elizabet,Elizavetinskij,Yelasavetinsk,Yelizabet,Yelizat,Yelizavet,Yelizavetinskiy,Елизаветинский	Russia, Sverdlovsk	section of populated place	N 56° 44' 49"	E 60° 36' 48"
6	Gornozavodskiy Gornozavodskij,Gornozavodskiy,Poselok Gornozavodskiy,Posëlok Gornozavodskiy,Горнозаводский	Russia, Sverdlovsk	section of populated place	N 56° 52' 0"	E 60° 35' 30"
7	Pyshma Pyshma,Пышма	Russia, Sverdlovsk	section of populated place	N 56° 56' 5"	E 60° 36' 49"
8	Novyy Novyj,Novyy,Новый	Russia, Sverdlovsk	section of populated place	N 56° 53' 51"	E 60° 38' 0"
9	Chermet Chermet,Чермет	Russia, Sverdlovsk	section of populated place	N 56° 46' 42"	E 60° 34' 58"
10	Moskovskiy	Russia,	section of	N 56° 48' 13"	E 60° 30' 9"

http://www.geonames.org/

Признаки

- Слова окружения (+/-)

директор, начальник, представитель…
компания, ООО, ПАО, фирма, …

- Части речи
  заявил/сказал/прибыл
- Структура

А. С. Пушкин, Михаил Лермонтов
Windows 10, iPhone 7

3.11.2016, (812) 212-85-06
Литейный пр-т, 24

NN4NER

- Lample et al., 2016
- две архитектуры:
  - LSTM-CRF*
  - Stack LSTM (~shift-reduce parsers)
- контекст, w2c + character-based represenatation
- 4 языка (Dutch, German, Spanish, English)
- результаты близкие или превосходящие SoA

Снятие неоднозначности

Иванов, Сергей

Материал из Википедии — свободной энциклопедии
(перенаправлено с «Сергей Иванов»)

Стабильная версия была проверена *12 августа 2016*. Имеются непроверенные изменения в шаблонах ил

В Википедии есть статьи о других людях с фамилией Иванов.

Серге́й Ивано́в:

- **Иванов, Сергей Александрович**:
  - Иванов, Сергей Александрович (1859—1927) — русский революционер, народоволец, заключённый Шлиссельбур
  - Иванов, Сергей Александрович (1870—1918/1919) — российский военно-морской деятель, контр-адмирал.
  - Иванов, Сергей Александрович
- **Иванов, Сергей Алексеевич**:
  - Иванов, Сергей Алексеевич (18
  - Иванов, Сергей Алексеевич (18
- **Иванов, Сергей Анатольевич**:
  - Иванов, Сергей Анатольевич (1
  - Иванов, Сергей Анатольевич (р
  - Иванов, Сергей Анатольевич —
  - Иванов, Сергей Анатольевич — администрации в 2005—2006 г
- **Иванов, Сергей Андреевич**:
  - Иванов, Сергей Андреевич (род. 1978) — российский игрок в мини-футбол.
  - Иванов, Сергей Андреевич (1922—1989) — Герой Советского Союза.

Задорнов

Материал из Википедии — свободной энциклопедии [править | править вики-текст]

Задо́рнов — русская фамилия. Известные носители:

- Задорнов, Михаил Михайлович (род. 1963) — российский политический деятель, в 1997—1999 — министр финансов РФ.
- Задорнов, Михаил Николаевич (род. 1948) — русский писатель-сатирик, сын Н. П. Задорнова.
- Задорнов, Николай Павлович (1909—1992) — русский писатель.

ИЗВЛЕЧЕНИЕ ОТНОШЕНИЙ (RELATION EXTRACTION)

20.10.2016, 11:59

Nissan приобрел контрольный пакет акций Mitsubishi Motors

Компания Nissan приобрела контрольный пакет акций японского концерна Mitsubishi Motors, заявил глава Nissan Motor и Renault Карлос Гон. Речь идет о 34% акций автопроизводителя. Господин Гон получил должность председателя совета директоров Mitsubishi Motors, в которую вступит с 14 декабря.

Помимо этого Nissan приобрел акции в Mitsubishi Corp., Mitsubishi Heavy Industries Ltd и Bank of Tokyo-Mitsubishi UFJ. Сообщается, что нынешний председатель совета директоров и президент Mitsubishi Motors Осама Масуко останется после завершения сделки на посту президента японской компании.

В рамках совместной работы Nissan и Mitsubishi намерены обмениваться технологиями, платформами и наработками по созданию электромобилей.

В текущем году компания Mitsubishi Motors впервые за восемь лет несет убытки из-за того, что в апреле призналась в фальсификации результатов тестов и искажении данных по экономичности двигателей автомобилей. Ожидается, что под руководством господина Гона Mitsubishi сможет поправить пошатнувшееся финансовое положение, а Nissan сможет расширить присутствие на тех рынках, где ранее популярностью пользовалась Mitsubishi.

http://kommersant.ru/doc/3120867

02.11.2016

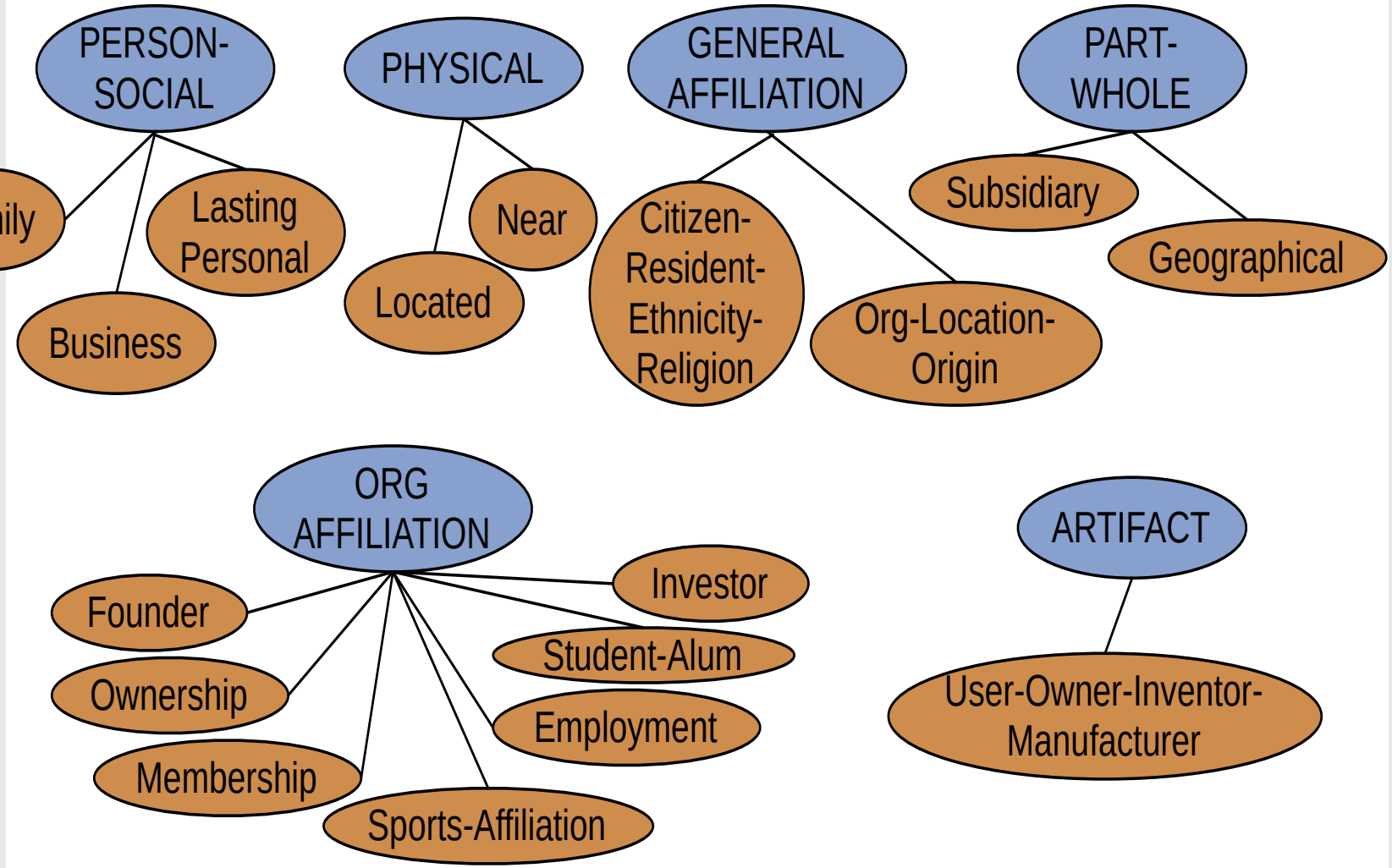
Владелец группы ЕСН Григорий Березкин и шведская инвесткомпания Kinnevik оказались не единственными учредителями московской версии Metro — самой популярной ежедневной газеты в стране. Совладельцами этого бизнеса стали наследник основателя "Балтийской медиагруппы" Олега Руднова Сергей и сын экс-гендиректора ЕСН Михаила Верозуба Владимир. Сейчас бизнес газеты реструктурируется, чтобы соответствовать требованиям закона "О СМИ".

Шведская Kinnevik вместе с родственниками и партнерами Григория Березкина учредила ООО "Газета номер один", следует из данных ЕГРЮЛ. 40,4% в этой компании получила сестра бизнесмена Ольга Березкина, 18,5% — его дочь Анна Тюшкевич, 1% — Kinnevik, 15% — сын бывшего гендиректора ЕСН Михаила Верозуба Владимир и 25,1% — ООО "Волна", контролирующееся структурами Сергея Руднова.

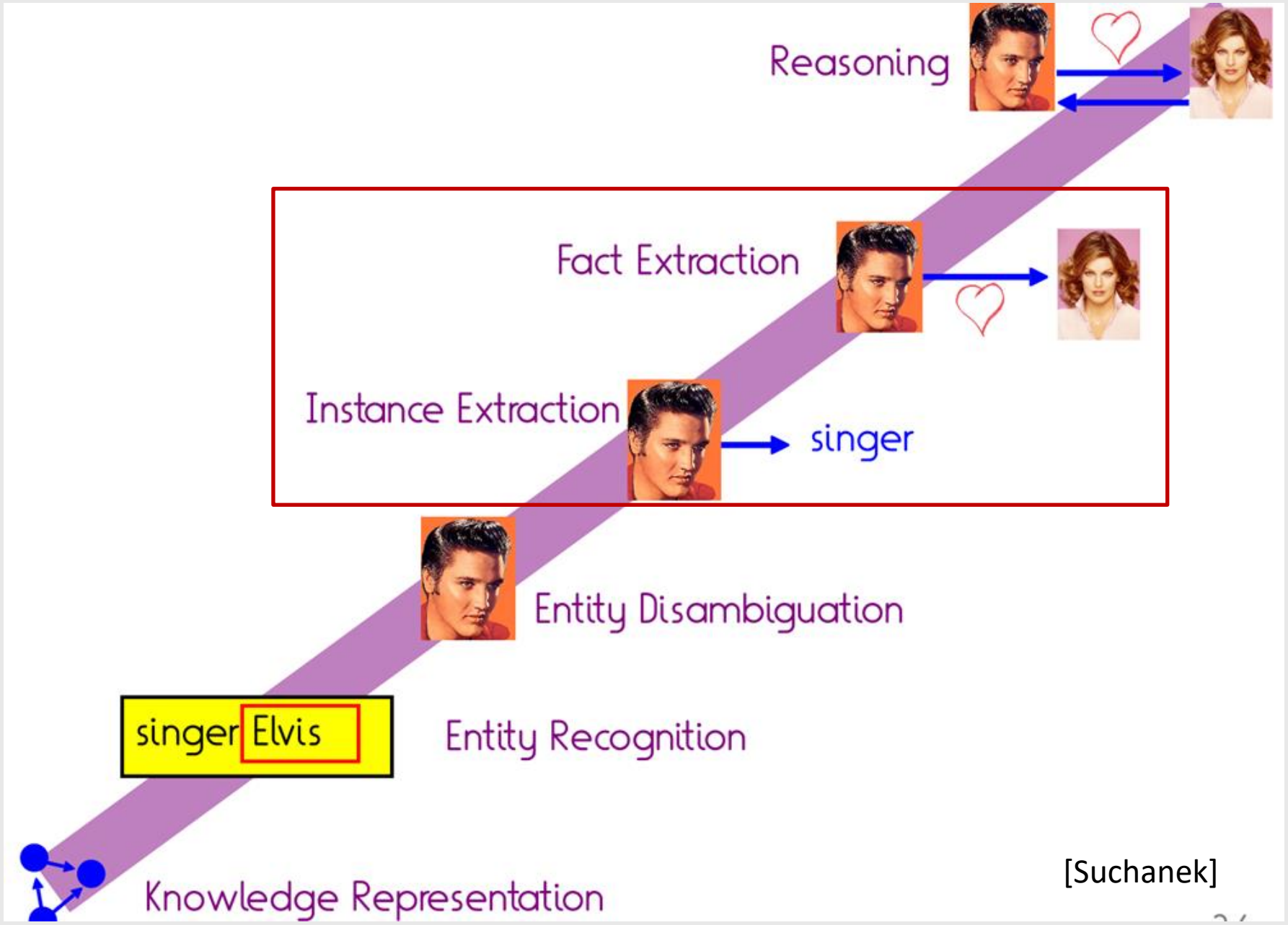
На ООО "Газета номер один" будет переведено управление газетой "Metro Москва", что связано с исполнением требований закона "О СМИ", пояснила "Ъ" представитель ЕСН Марианна Белоусова. Подробности она не уточнила, отметив, что существенных акционерных изменений в газете нет. В самой газете отказались комментировать эту информацию. Свидетельство о СМИ "Газета Metro Москва" до сих пор оформлено на старую структуру — АО "Газета "Метро"", следует из реестра Роскомнадзора. В ЕГРЮЛ актуальные акционеры этой компании недоступны. Ранее по 9,5% в газете принадлежало департаменту имущества Москвы и ГУП "Московский метрополитен". В пресс-службе ГУП сообщили, что предприятие сохраняет долю в газете. Представитель департамента городского имущества и представитель департамента рекламы и СМИ Москвы говорят, что сейчас на балансе города акций газеты нет.

http://kommersant.ru/doc/3132404

Automated Content Extraction (ACE)



[Dan Jurafsky]


Reasoning
Fact Extraction
Instance Extraction
singer
Entity Disambiguation
singer Elvis
Entity Recognition
Knowledge Representation

[Suchanek]

Методы

- Шаблоны
- Классификация «с учителем» (supervised)
- С частичным привлечением учителя (semi-supervised) и без учителя (unsupervised)
  - bootstrapping (использование затравки – seeds)
  - «удаленное обучение» (distant supervision)
  - обучение без учителя

Выделение гиперонимов (Hearst, 1992)

Pattern	Example
X and other Y	...temples, treasuries, and other important civic buildings.
X or other Y	Bruises, wounds, broken bones or other injuries...
Y such as X	The bow lute, such as the Bambara ndang...
Such Y as X	...such authors as Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, including Canada and England...
Y , especially X	European countries, especially France, England, and Spain...

[Dan Jurafsky]

Словарные определения

Э́лвис Аро́н Пре́сли (англ. *Elvis Aron Presley*[3]; 8 января 1935, Ту́пело — 17 августа 1977, Мемфис) — американский певец и актёр, один из самых коммерчески успешных исполнителей популярной музыки XX века[4]. В Америке Пресли прозвали «королём рок-н-ролла» (или просто «королём» — *The King*). Он популяризовал рок-н-ролл, хотя не был первым исполнителем этого жанра. Тем не менее, именно он соединил кантри и блюз, дав рождение рокабилли[4], которым отмечены его первые записи на Sun Records в середине 1950-х гг. Вкрапливая в свой стиль

Дима Никола́евич Била́н[1][2] (имя при рождении и до июня 2008 года — **Ви́ктор Никола́евич Бела́н**; род. 24 декабря 1981, пос. Московский, часть города Усть-Джегута[2], Карачаево-Черкесская АО, РСФСР, СССР) — российский певец и киноактёр. В июне 2008 года принял данный псевдоним в качестве настоящего имени и фамилии[3]. Заслуженный артист Кабардино-Балкарии (2006), Заслуженный артист Чечни (2007), Заслуженный артист Ингушетии (2007) и Народный артист Кабардино-Балкарии (2008).

Правила + NEs

PERSON (f) вышла замуж за PERSON (m)

PERSON (m) женился на PERSON (f)

директор ORG PERSON

PERSON назначен директором ORG

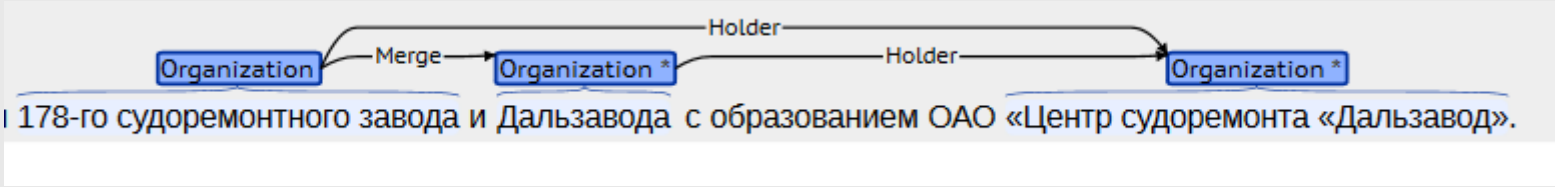
Правила, составленные вручную

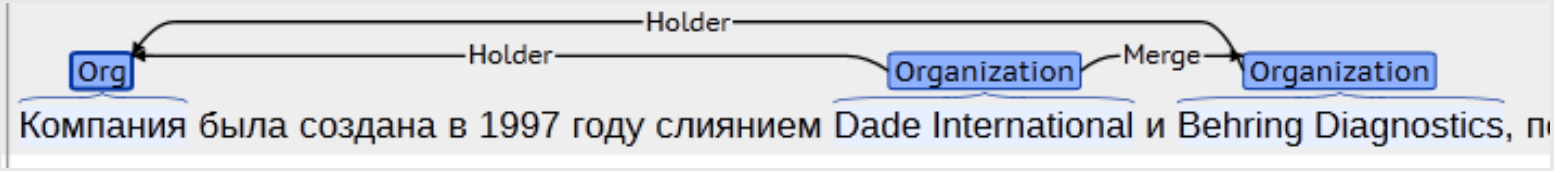
- +
  - высокая точность
  - специализация
- –
  - низкая полнота
  - много ручной работы
  - плохая переносимость

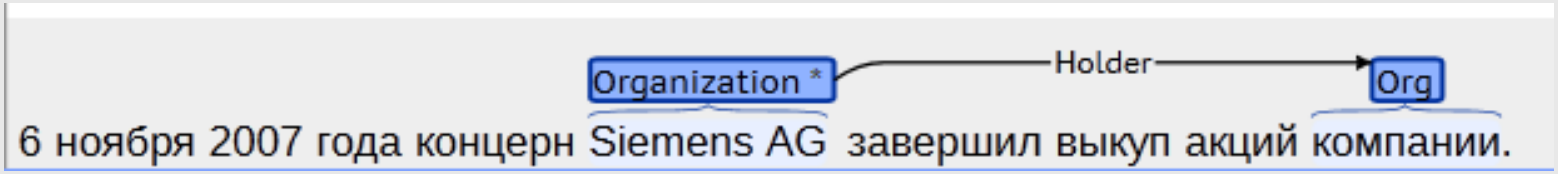
Машинное обучение

- Разметить данные (NEs, отношения)
- Разработать набор признаков
- Обучить классификатор
- Оценить качество

Разметка







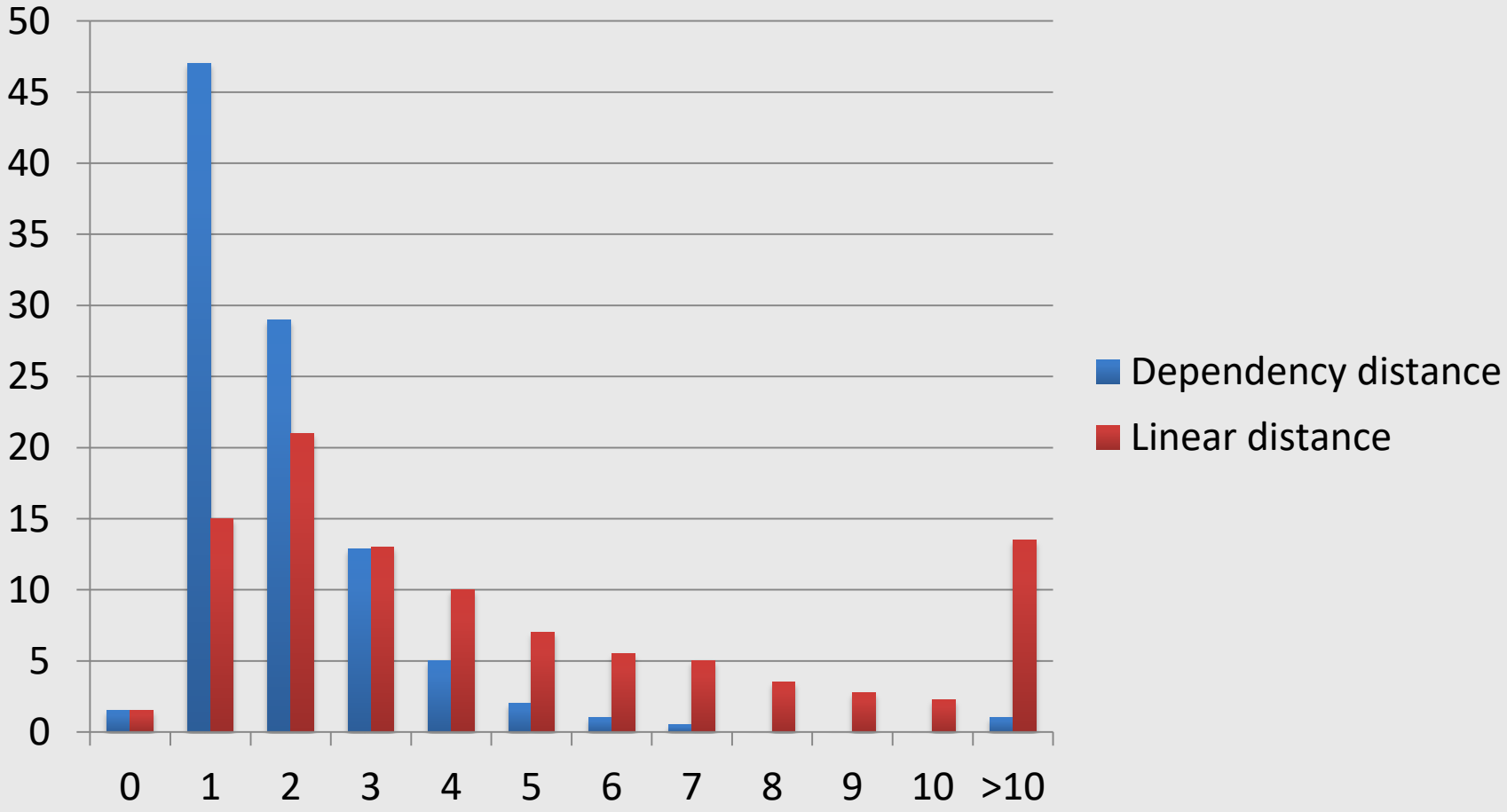
http://brat.nlplab.org/

***American Airlines**, a unit of AMR, immediately matched the move, spokesman **Tim Wagner** said.*

Entity-based features	
Entity_1 type	ORG
Entity_1 head	*airlines*
Entity_2 type	PERS
Entity_2 head	*Wagner*
Concatenated types	ORGPERS
Word-based features	
Between-entity bag of words	{ *a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman* }
Word(s) before Entity_1	NONE
Word(s) after Entity_2	*said*
Syntactic features	
Constituent path	$NP \uparrow NP \uparrow S \uparrow S \downarrow NP$
Base syntactic chunk path	$NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP$
Typed-dependency path	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

[Dan Jurafsky]

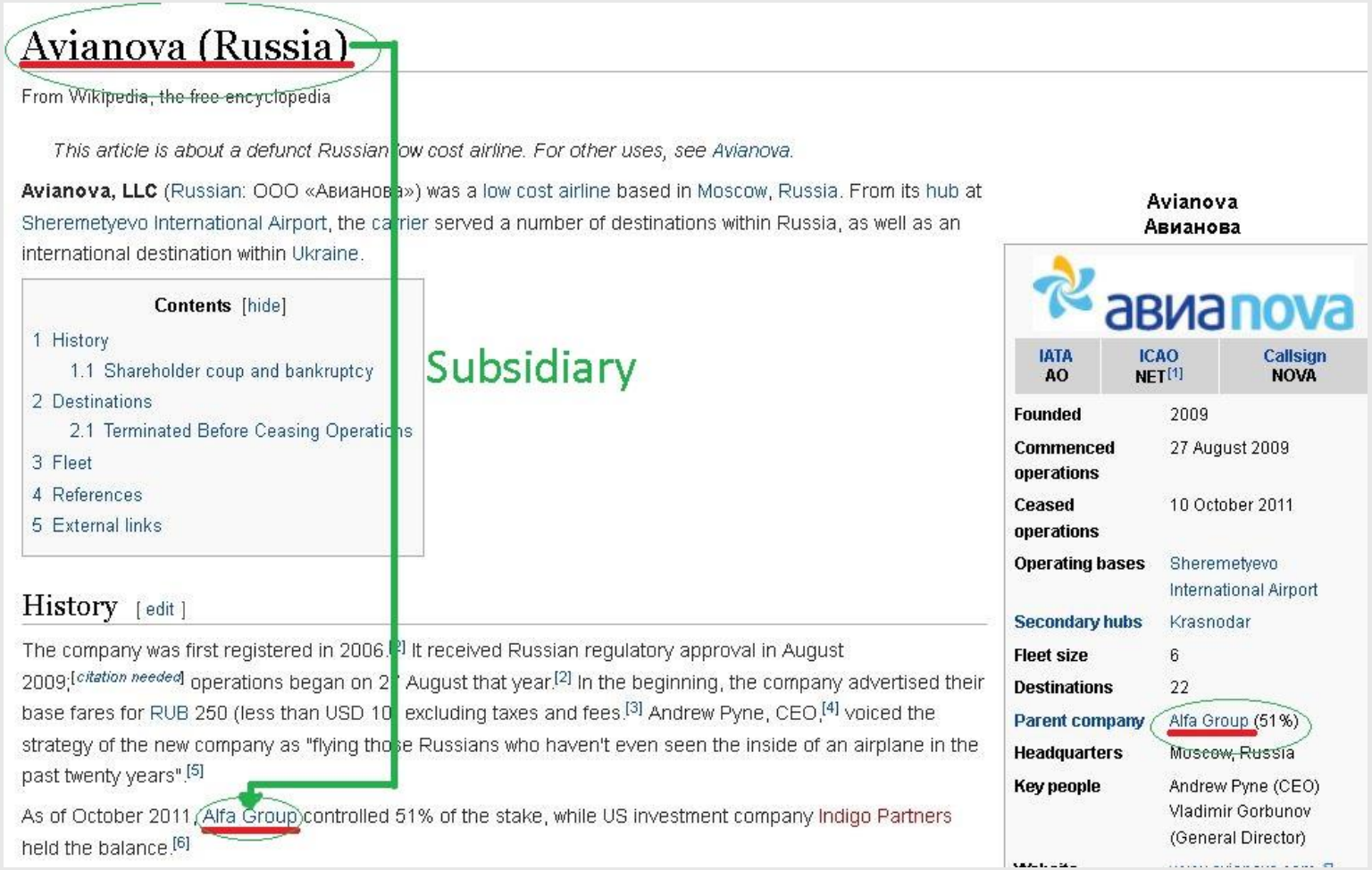
BioNLP 2009/2011 relation extraction



[Björne et al. 2009]

[Chris Manning]

Автоматическая разметка



Bootstrapping / затравка (seeds)

- Начинаем с небольшого набора образцов отношений или высокоточного шаблона

  СУПРУГИ (Александр Пушкин, Наталья Гончарова)
- Находим упоминания, генерируем шаблоны, повторяем

Примеры

8 декабря 1863 года умерла Наталья Гончарова, супруга Александра Сергеевича Пушкина.

…из-за конфликтов с будущей тёщей и отсутствия денег с обеих сторон пожениться Гончарова и Пушкин смогли лишь спустя три года.

В 1831 году Пушкин женится на красавице Н. Н. Гончаровой…

Гончарова вышла замуж за Пушкина 19-ти лет; поэту было тогда 32 года.

2 марта 1831 года, в Москве, в церкви Большого Вознесения на Большой Никитской, был совершен обряд венчания А.С. Пушкина с Н.Н. Гончаровой.

«Удаленное обучение»

- bootstrapping + обучение с учителем
  - очень много примеров отношений
  - большое количество признаков
  - построение классификатора
- «обучение с учителем»
  - классификатор с большим набором признаков
  - обучающая информация в виде примеров
  - не требует итеративного расширения шаблонов
- «обучение без учителя»
  - использует большие коллекции неразмеченных данных
  - лучше справляется с жанровым разнообразием

Snow, Jurafsky, Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. NIPS 17
Fei Wu and Daniel S. Weld. 2007. Autonomously Semantifying Wikipeida. CIKM 2007
Mintz, Bills, Snow, Jurafsky. 2009. Distant supervision for relation extraction without labeled data. ACL09

[Dan Jurafsky]

«Удаленное обучение»

1. Отношения
2. Примеры в базе
3. Предложения с обоими участниками
4. Частотные признаки (синтаксические отношения, слова, …)
5. Классификатор на основе тысяч шаблонов

Born-In

<Edwin Hubble, Marshfield>
<Albert Einstein, Ulm>

Hubble was born in Marshfield
Einstein, born (1879), Ulm
Hubble's birthplace in Marshfield

PER was born in LOC
PER, born (XXXX), LOC
PER's birthplace in LOC

$$P(\text{born-in} \mid f_1, f_2, f_3, \ldots, f_{70000})$$

[Dan Jurafsky]

Извлечение «без учителя»

Идея: извлечь все отношения из большого корпуса

1. На основе размеченных данных строится классификатор для троек (*trustworthiness*)
2. Извлекаются все отношения между именными словосочетаниями, отфильтровываются «ненадежные»
3. Оценка вероятности отношения на основе избыточности (повторов)

(FCI, specializes in, software development)

(Tesla, invented, coil transformer)

M. Banko, M. Cararella, S. Soderland, M. Broadhead, and O. Etzioni. 2007. Open information extraction from the web. IJCAI

Open Information Extraction

http://openie.allenai.org/

Hosted by AI2

Argument 1: smoking **Relation:** causes **Argument 2:** All Search

74 answers from **1611 sentences** (cached)

all | disease (25) | cause of death (22) | risk factor (14) | medical condition in fiction (11) | icd 9 cm classification (10) | misc. | more types

- **Lung cancer** (460)
- **Cancer** (280)
- **Heart disease** (113)
- **Chronic obstructive pulmonary disease** (46)
- **Emphysema** (27)
- **Erectile dysfunction** (19)
- **Halitosis** (14)
- **Respiratory disease** (12)
- **Asthma** (9)
- **Bladder cancer** (8)
- **Insomnia** (7)
- **Cough** (7)
- **Myocardial infarction** (6)
- **Cardiovascular disease** (6)
- **Stroke** (5)
- **Chronic bronchitis** (5)

Lung cancer

URI:

http://www.freebase.com/view/m/04p3w

Types:

disease (Nell)
/medicine/disease (FreeBase)
/people/cause_of_death (FreeBase)
/fictional_universe/medical_condition_in_fiction (FreeBase)

Extracted Synonyms:

heart disease and lung cancer
lung
his lung cancer
most lung cancers
the lung cancer
lung cancers
most lung cancer

Extracted from these sentences:

causes

Cigarette smoking causes **lung cancer** , heart attacks , strokes , emphysema and other diseases . (via ClueWeb12)

Cigarette smoking causes **lung cancer** , heart disease , emphysema , and other serious diseases in smokers . (via ClueWeb12)

Cigarette smoke causes **lung cancer** , heart disease and a host of other diseases . (via ClueWeb12)

smoking causes **lung cancer** , heart diseases , emphysema , and may complicate relationship . (via ClueWeb12)

That smoking causes **lung cancer** , heart disease , emphysema , and other diseases is universally accepted by medical and scientific authorities . (via ClueWeb12)

Secondhand smoke causes **lung cancer** , other cancer , heart disease and both major a

ИНСТРУМЕНТЫ

Инструменты

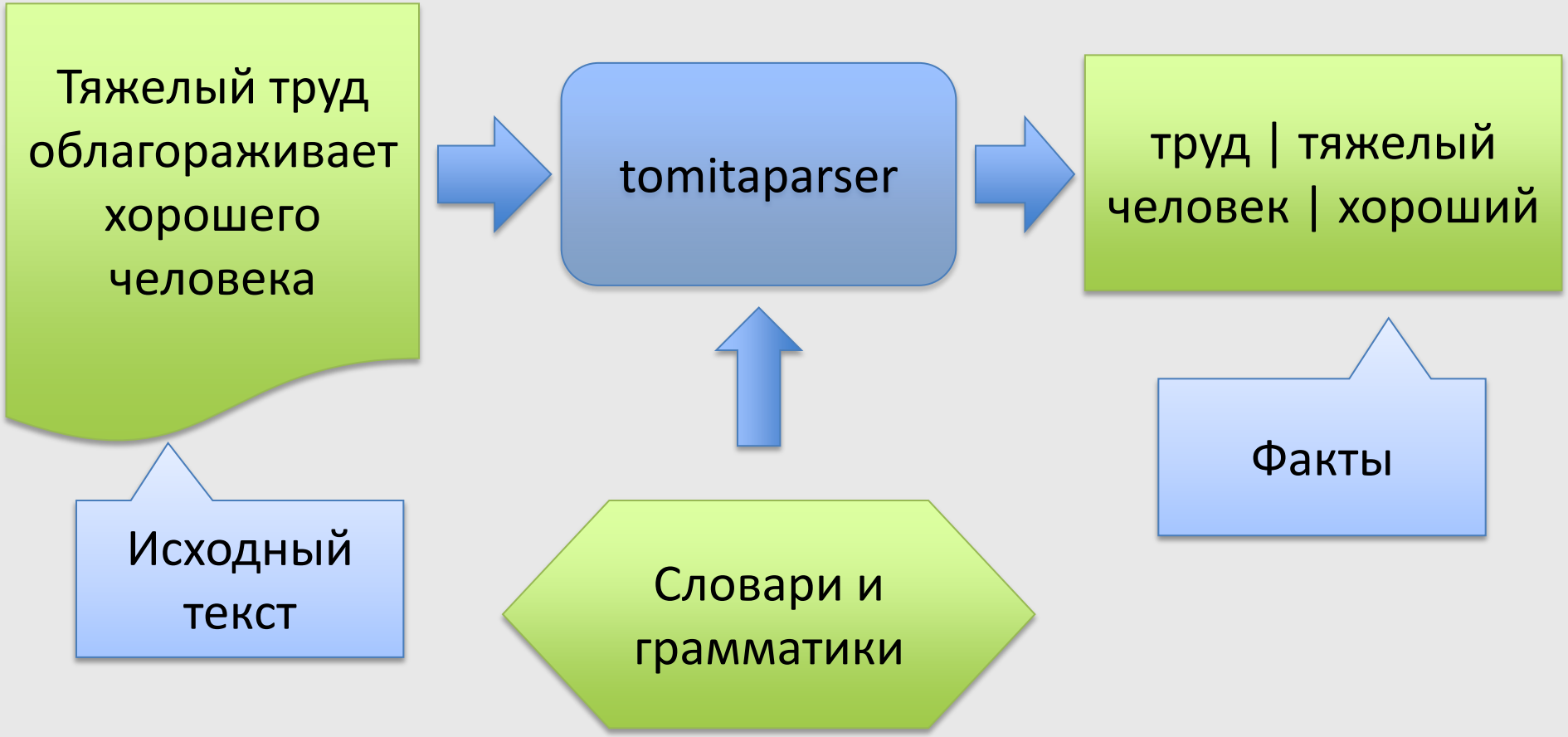
- регулярные выражения
- словари
- классификаторы
- обучение на последовательностях (CRF, HMM)
- обучение правил (Apache UIMA TextRuler)
- инструменты для составления шаблонов (Tomita)

Томита-парсер

- Составленные вручную правила
  - контекстно-свободная грамматика
    S → NP VP
    NP → Noun
    NP → Adj NP
- лингвистические признаки (POS, граммемы, согласование) + словари + регулярные выражения
- алгоритм GLR – парсинга (http://ru.wikipedia.org/wiki/GLR-парсер)
- назван в честь японского ученого Масару Томита, автора алгоритма
- Используется в Яндексе (Новости, Работа)
- Проект с открытым кодом

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

Что делает Томита-парсер?



[Ландо, Остапук]

Пример: имена

```
Initial -> Word<wff=/[А-Я]\./>;
Initials -> Initial Initial;
Person -> Initials Word<h-reg1>;
Person -> Word<kwtype=first_names,h-reg1,gnc-agr[1]>
Word<h-reg1,gnc-agr[1]>;
```

ОЦЕНКА/ДАННЫЕ

Оценка

- Границы/тип
  Действующий чемпион мира по шахматам норвежец Магнус Карлсен, который после поражения от россиянина…
- Неоднозначности
  Берлин озабочен проблемой беженцев.
  Презентация прошла в магазине Библио-Глобус.

Message Understanding Conference (MUC)

Conference	*Year*	*Text Source*	*Topic (Domain)*
MUC-1	1987	Mil. reports	Fleet Operations
MUC-2	1989	Mil. reports	Fleet Operations
MUC-3	1991	News reports	Terrorist activities in Latin America
MUC-4	1992	News reports	Terrorist activities in Latin America
MUC-5	1993	News reports	Corporate Joint Ventures, Microelectronic production
MUC-6	1995	News reports	Negotiation of Labor Disputes and Corporate Management Succession
MUC-7	1997	News reports	Airplane crashes, and Rocket/Missile Launches

Automated Content Extraction (ACE)

ACE 2008

ACE 2008 tasks included Local (within-document) EDR (Entity Detection and Recognition) and RDR (Relation Detection and Recognition) for English and Arabic. ACE 2008 also included a pilot task for Global (cross-document) EDR and RDR for English and Arabic.

- English Entities V6.6
- English Relations V6.2
- Arabic Entities V7.4.2
- Arabic Relations V6.5
- Cross-Document Coreference

ACE 2007

Evaluation

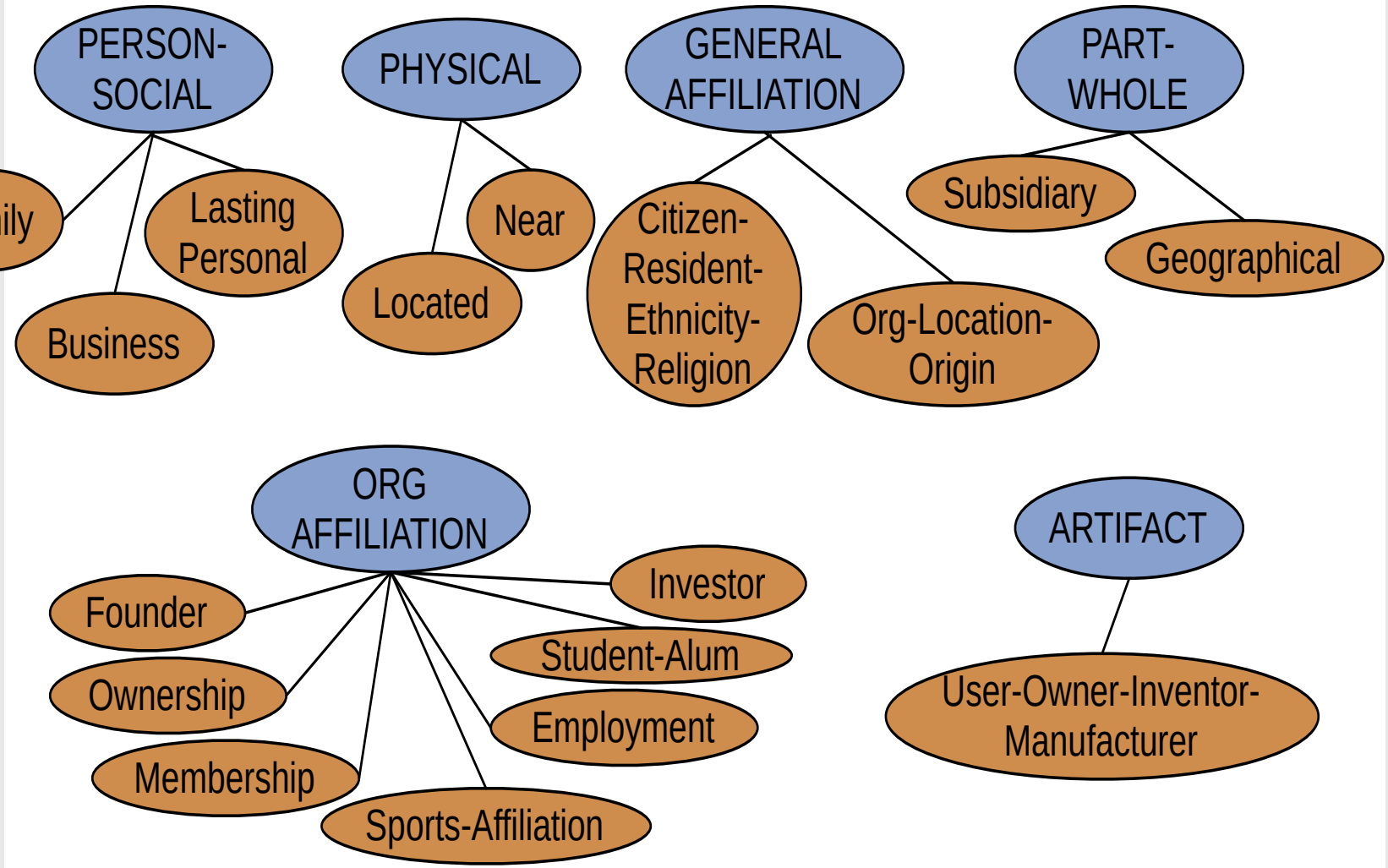
Tasks for ACE 2007 included a pilot evaluation using Spanish data for entity detection and recognition (EDR) and temporal expression recognition and normalization (TERN). Data selection was semi-automatic; a set of candidate documents was manually reviewed to select individual documents that were suitable for ACE annotation, for instance documents that were representative of their genre and contain targeted ACE entity types.

- Spanish Entities V1.6

Entity Translation Pilot Evaluation

A pilot evaluation of Entity Translation was also conducted as part of ACE 2007. Systems participating in the pilot ET task were evaluated on their ability to take in a text document in one

Automated Content Extraction (ACE)



[Dan Jurafsky]

Text Analysis Conference (TAC)

TAC Track Home Pages									
All Tracks	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
Question Answering	2008								
Recognizing Textual Entailment	2008	2009	2010	2011					
Summarization	2008	2009	2010	2011			2014		
Knowledge Base Population		2009	2010	2011	2012	2013	2014	2015	2016

РОМИП

- 2004:
  - по описанию персоны найти в Веб-коллекции все связанные факты
  - ответ: фрагмент документа + тип факта (опционально)
- 2005:
  - именованные сущности из новостей
    - персона, организация, геообъект, прочее
  - факты:
    - работает (персона, организация)
    - владеет (персона, организация)
    - владеет (организация, организация)

http://www.romip.ru/ru/2005/tracks/qa.html

Результаты оценки (2005)

- оценка на основе ответов систем (а не предварительно подготовленный «золотой стандарт»)
- ~400 документов → ~2600 фактов

```
<fact  firstText="ОТДЕЛ ГОСУДАРСТВЕННОГО УПРАВЛЕНИЯ ОХРАНОЙ ТРУДА"
secondText="БАКИН АЛЕКСАНДР" systemLength="318" userLength="2" systemOffset="110"
userOffset="18" category="employs" entityOffset="83" collectionIdRef="ROMIP-2005-
News" relevance="vital"/>

<fact  firstText="ГАЗПРОМ" secondText="LIETUVOS DUJOS" systemLength="185"
userLength="69" systemOffset="50035" userOffset="50058" category="owns"
entityOffset="33" collectionIdRef="ROMIP-2005-News" relevance="vital"/>

<fact  firstText="РОССИЯ" secondText="Совершаева Любовь" systemLength="77"
userLength="77" systemOffset="11590" userOffset="11590" cathegory="employs"
entityOffset="50" collectionIdRef="ROMIP-2005-News" relevance="notrelevant"/>
```

http://romip.ru/relevance-tables/ru/index.html

Данные «Связи компаний» (2016)

Источник: русская Википедия, страницы организаций

Отношение: владение

2,150 Holder/ 992 Subsidiary/4,012 Other

```
Загребская пивоварня
…
Расположена в столице страны Загребе, принадлежат активам
международной пивоваренной корпорации StarBev.
```

```
T1      Organization 0 20      Загребская пивоварня
T2      Organization 233 240   StarBev
N1      Reference T1 wiki:4114742     Загребская пивоварня
N2      Reference T2 wiki:4114777     StarBev
A1      IsLink T2
A2 TypeRel T2 Holder
```

http://kansas.ru/pb/paper/kesw2016.pdf

Диалог 2016

- Именованные сущности:
  - Персоны
  - Организации
  - Локации
- Именованные сущности: атрибуты, нормализация, идентификация
- Факты (отношения)
  - Работает (персона, организация)
  - Сделка (участник1, участник2, …)
  - Владеет (персона, организация)
  - Встреча(персона1, персона2, …)

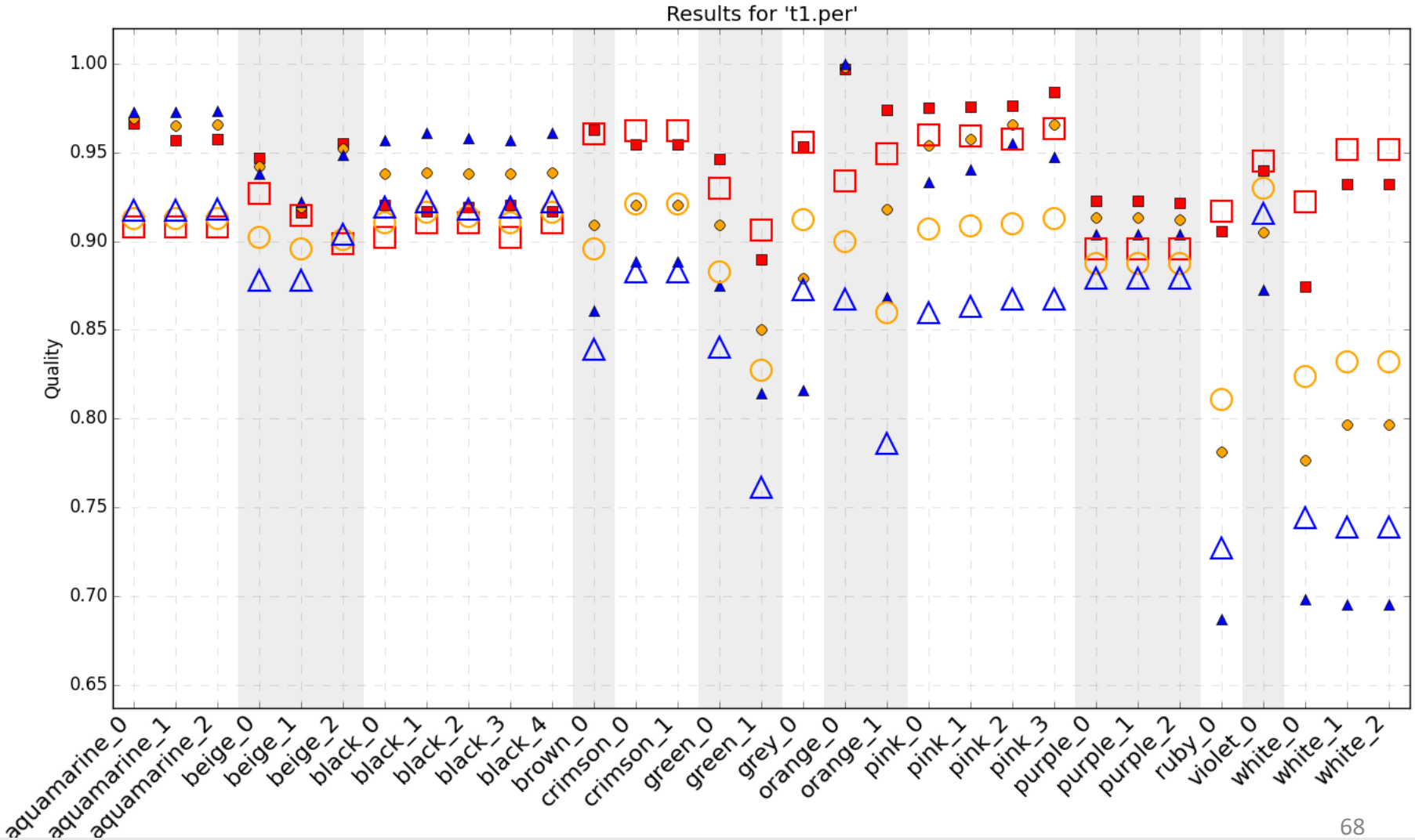
http://www.dialog-21.ru/evaluation/2016/ner/

Данные

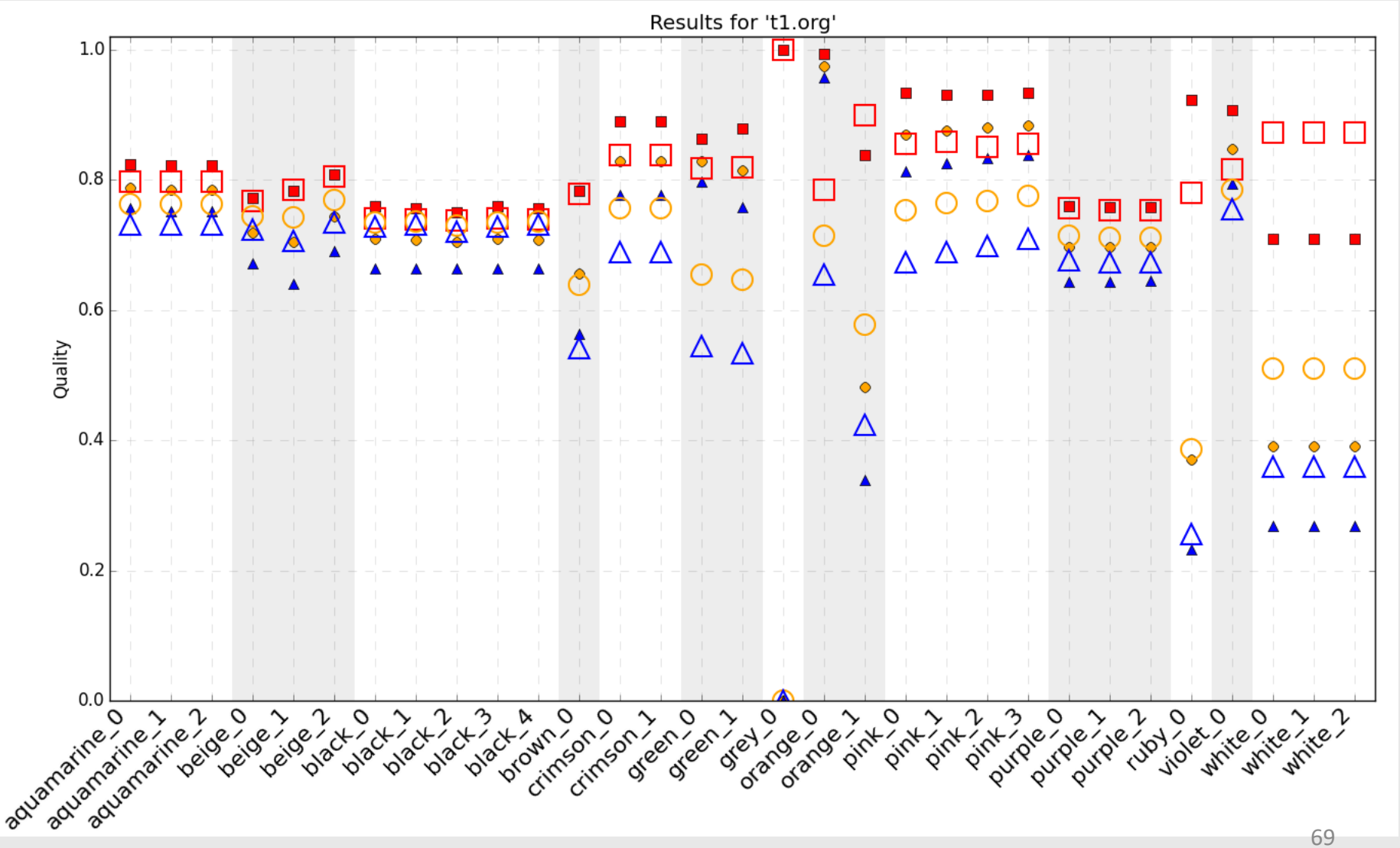
- ~250 новостных сообщений (демо/тест)
- 13 участников
- объем разметки:

Objects	
Demo	Test
2,611	5,019
Facts	
Demo	Test
273	786

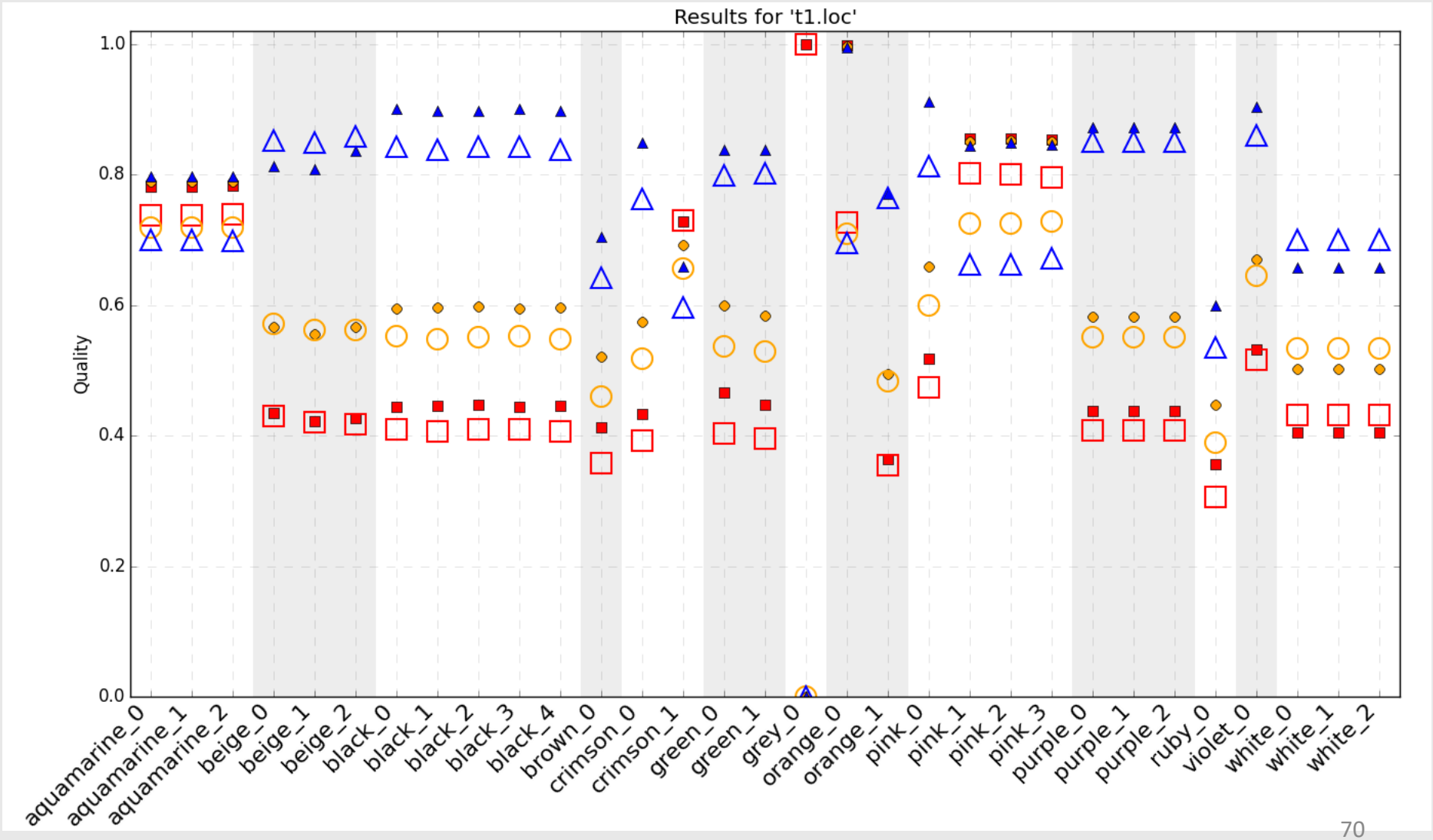
Выделение персон



Выделение организаций



Выделение локаций



Выделение фактов