

Проект

Извлечение фактов

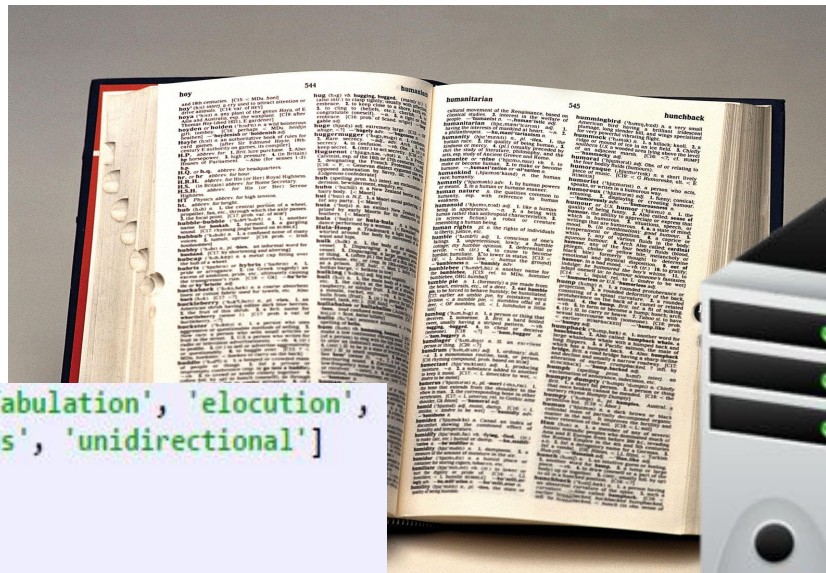
из текста

Лаборатория

математической лингвистики

Что такое компьютерная лингвистика?

Компьютерная лингвистика изучает язык с позиции его использования в компьютерных системах.



```
>>> words = ['attribution', 'confabulation', 'elocution',  
...          'sequoia', 'tenacious', 'unidirectional']  
>>> vsequences = set()  
>>> for word in words:  
...     vowels = []  
...     for char in word:  
...         if char in 'aeiou':  
...             vowels.append(char)  
...     vsequences.add(''.join(vowels))  
>>> sorted(vsequences)  
['aiuo', 'eaio', 'eouio', 'euoia', 'ouaio', 'uieioa']
```



Задачи компьютерной ЛИНГВИСТИКИ:

- автоматическое составление словарей и грамматик;
- анализ естественно-языковых текстов;
- создание и использование текстовых корпусов;
- машинный перевод;
- информационный поиск;
- автореферирование;
- создание систем искусственного интеллекта и др.



- Извлечение фактов (структурированной информации) из неструктурированного текста - **Text Mining**.
- С помощью этой технологии можно представлять данные из текстов на естественном языке в формализованном виде для дальнейшей **машинной обработки**.
- Извлечение фактов - одна из задач **компьютерной лингвистики**.

Где применяются технологии извлечения фактов?

Яндекс – Почта, Новости, Карты и др. сервисы.

Встреча



Анна Кудрявцева



Тема: встреча

Перенести Сдать заявку Сделать запись

Детский день!

Место встречи - [г. Новосибирск, Морской пр-т, 54](#) Увидимся 20 августа!



Где применяются технологии извлечения фактов?

Встреча



Anna Kozlova anna2016@gmail.com

Контакт anna2016@yandex.ru

[Перевести](#) [Создать правило](#) [Свойства письма](#)

Добрый день!

Место встречи - [г.Новосибирск, Морской пр-т, 54](#). Увидимся 20 августа!

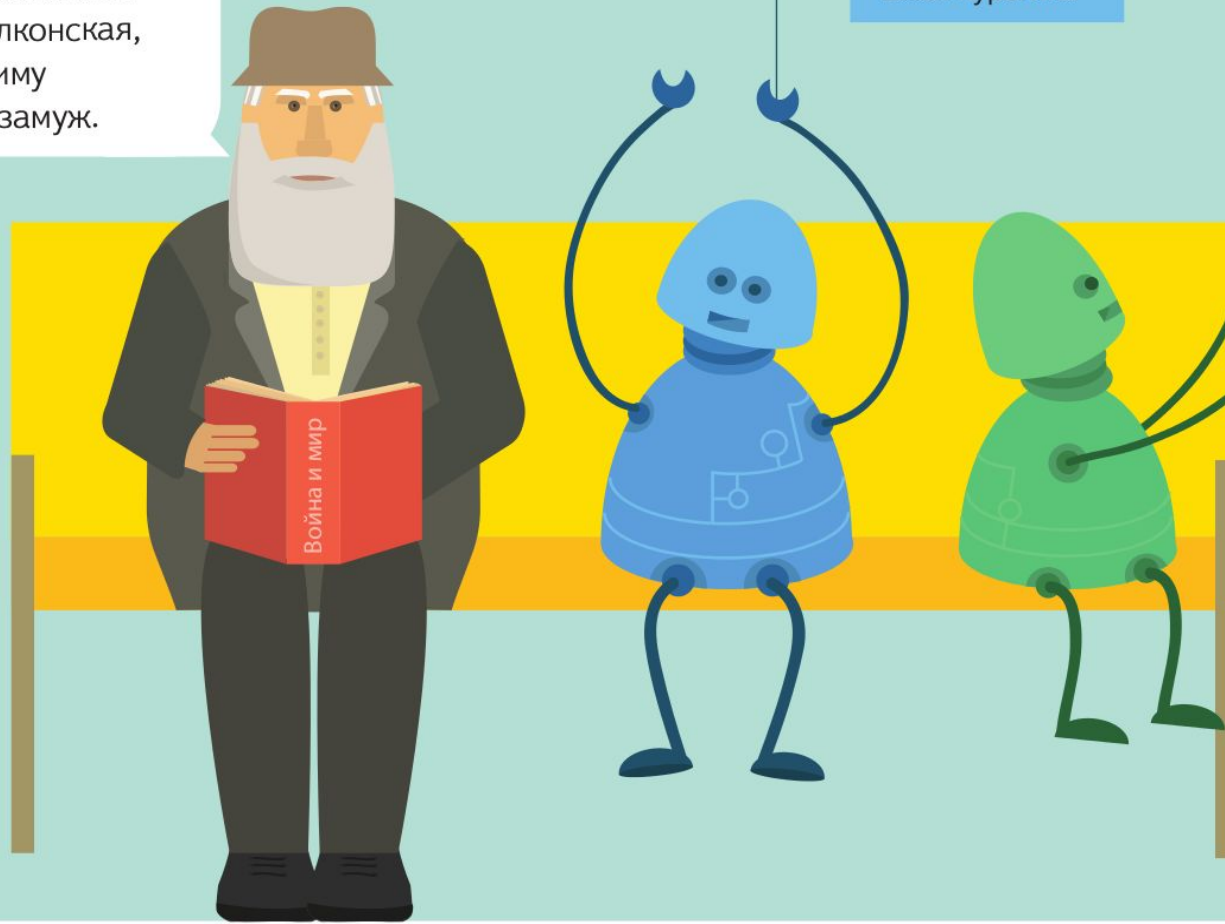
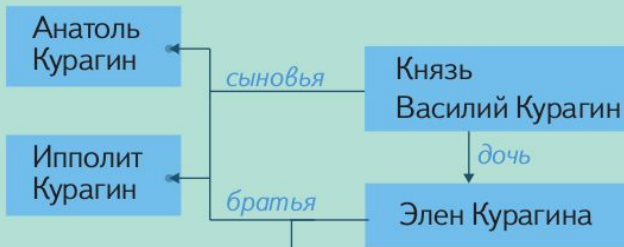


Где применяются технологии извлечения фактов?

- В поисковых системах, например Google и Yandex, для сбора информации о пользователе.
- При автоматическом построении предметных областей.
- Для представления текстовой информации в удобном виде для машинной обработки.

Пример извлечения фактов

Вошла дочь князя Василия, красавица Элен, захавшая за отцом, чтобы с ним вместе ехать на праздник посланника. Приехала и известная как самая обворожительная женщина Петербурга молодая, маленькая княгиня Болконская, прошлую зиму вышедшая замуж.



Задача проекта:

извлечение фактов из текстов для структурирования информации.

Под «фактом» понимается набор извлеченных сущностей, связанных определенным отношением.

Источник: научные тексты по химии.

Примеры неструктурированного текста:

- В 1771 году **Карл Шееле** получил плавиковую кислоту.
- В природе значимые скопления **фтора** содержатся в основном в минерале **флюорите** (**CaF₂**).
- **Глюкоза** - бесцветное **кристаллическое вещество** сладкого вкуса, **растворимое в воде**.
- При окислении образует **глюконовую кислоту**.

Получаем на выходе:

Название вещества	Формула	Физические свойства	Температура плавления	Температура кипения
глюкоза	$C_6H_{12}O_6$	кристаллическое вещество сладкого вкуса, растворимое в воде	150 °C	
гидроксид меди(I)	$CuOH$	жёлтое вещество, не растворяется в воде		
аммиак	NH_3	бесцветный газ с резким запахом	-77.73 °C	-33.34 °C

Инструменты для работы

- **Томи́та-парсер** — это инструмент для извлечения структурированных данных (фактов) из текста на естественном языке. Это технология, разработанная Яндексом.
- Для извлечения информации из текста с помощью томи́та-парсера нужно писать **грамматики**.

Грамматика томига-парсера

Так выглядит часть грамматики для томига-парсера (для извлечения места рождения человека):

Born -> Verb<kwtype=born>;

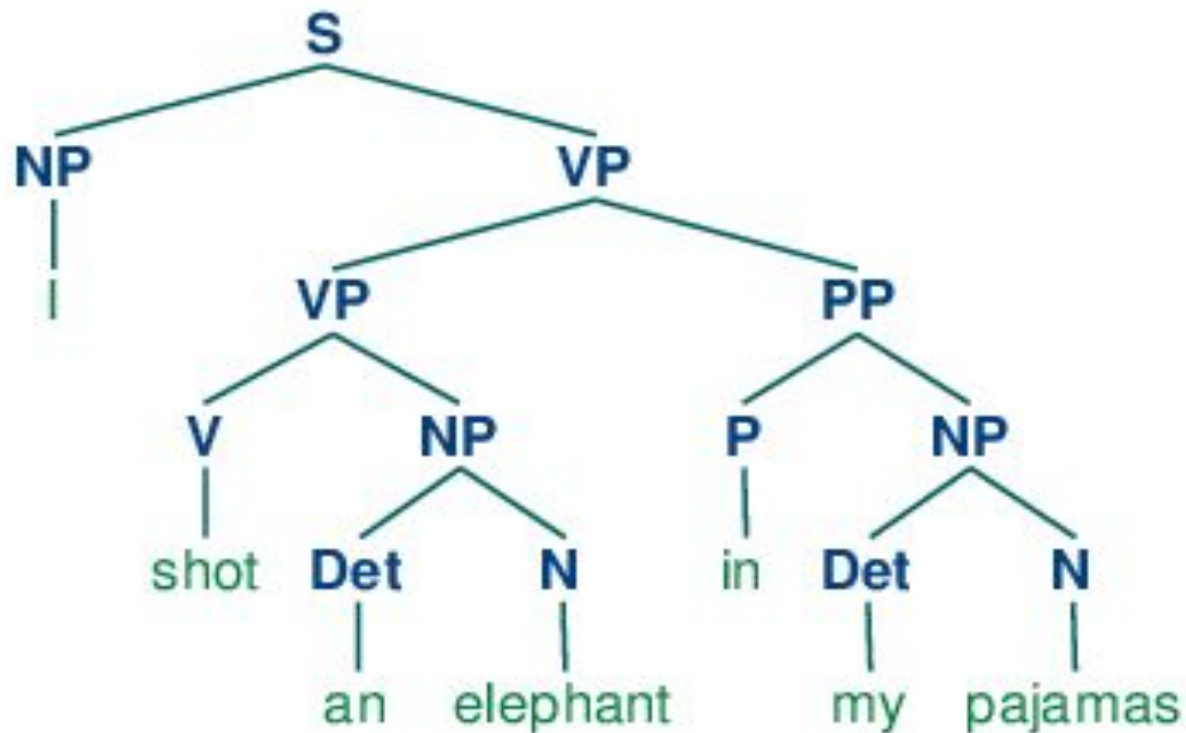
City -> Noun<kwtype=city>;

Person -> AnyWord<gram="имя">;

S -> Person interp(BornFact.Person) Born "в"
City interp(BornFact.Place);

Грамматика томита-парсера

- Язык описания грамматик для томита-парсера построен на основе **порождающих грамматик**.



ИСТОЧНИКИ:

- Блог Яндекса на Хабре
<http://habrahabr.ru/company/yandex/blog/219311/>
- <http://habrahabr.ru/company/yandex/blog/205198/>
- Скриншоты с Яндекс Почты

Спасибо за внимание!