

Компьютерная лингвистика

Введение

Задачи АОТ. Этапы обработки

- исторически 3 разных понимания термина «компьютерная лингвистика»:
 - инструментарий для обработки лингвистических данных
 - формальные модели (Computational linguistics) / вычислительная лингвистика;
 - современная автоматическая обработка естественного языка (Natural Language Processing)

- **Инженерная компьютерная лингвистика**
- междисциплинарная область, в задачи которой входит автоматический анализ текстов: **автоматическая обработка ЕЯ**
- **Инструментальная компьютерная лингвистика.**
компьютерные технологии для обработки текстов, для представления лингвистических данных
(корпуса, лингвистические ресурсы, парсеры).
- **Теоретическая компьютерная лингвистика (вычислительная лингвистика):**
применение математических (формальных) моделей к описанию естественного языка, моделирование функционирования языка с использованием формального аппарата.

Общий взгляд на задачи компьютерной лингвистики

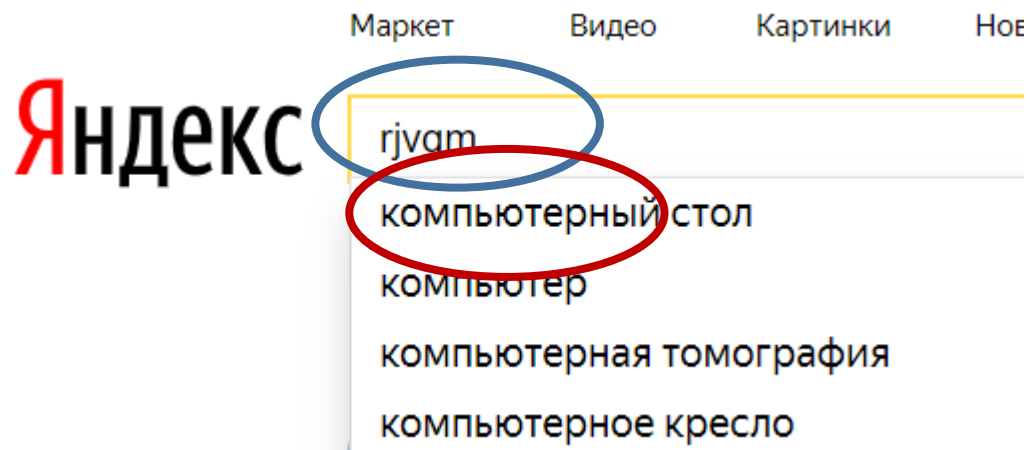
- Инженерная компьютерная лингвистика
 - междисциплинарная область, в задачи которой входит автоматический анализ текстов: автоматическая обработка ЕЯ
- Инструментальная компьютерная лингвистика.
 - компьютерные технологии для обработки текстов, для представления лингвистических данных
(корпуса, лингвистические ресурсы, парсеры).
- Теоретическая компьютерная лингвистика (вычислительная лингвистика):
 - применение математических (формальных) моделей к описанию естественного языка, моделирование функционирования языка с использованием формального аппарата.

Вопросы к обсуждению

- С какими приложениями, использующими автоматический анализ текста, мы сталкиваемся каждый день?
- <https://yandex.ru/>

Примеры задач

- распознавание языка



Примеры задач

- опечатки

Яндекс

маркет видео картинки новости

кампю|

компьютерный стол

компьютер

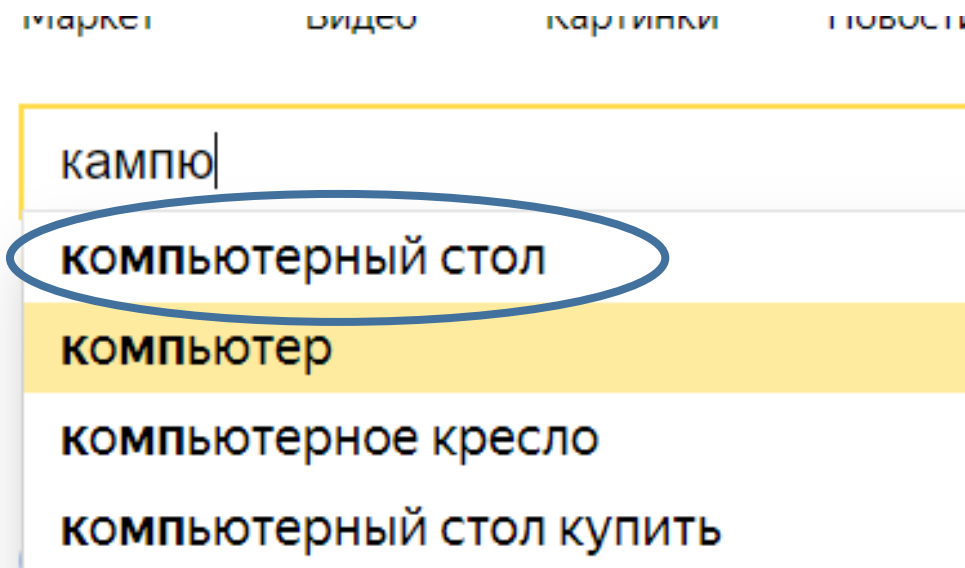
компьютерное кресло

компьютерный стол купить

Примеры задач

- устойчивые словосочетания

Яндекс



Примеры задач

- автоматическое распознавание событий, реферирование





14.10, 16 20 ⋮

- iz** Путин объявил о регистрации второй российской вакцины от коронавируса
- В** Россия возобновляет авиасообщение с Сербией, Кубой и Японией
- Г** Послы Евросоюза согласовали санкционный «список Навального»
- L** Навальный раскрыл имена оплативших его лечение в Германии людей
- iz** Московским школьникам 1-5 классов разблокируют транспортные карты

Примеры задач

- группировка текстов по событиям
- <https://yandex.ru/promo/news/index>
- Из полученных материалов сервис формирует сюжеты — подборки новостных сообщений об одном событии, которые помогают ориентироваться в потоке новостей

Россия с 1 ноября возобновит авиасообщение с Японией

-  Россия возобновляет авиасообщение еще с тремя странами
Самолеты начнут летать в Сербию, Кубу и Японию. ...Согласно распоряжению, подписанному премьером Михаилом Мишустиним, в Сербию и на Кубу их будет по два в неделю. **Эхо Москвы 16:06**
-  Россия возобновит авиасообщение с несколькими странами
В этот список вошли Сербия, Куба и Япония, сообщает РИА Новости. По данным агентства, самолет из Москвы в Белград и на Кубу будут летать два раза в неделю, а из Москвы в Японию - один раз в неделю. **АМИТЕЛ 16:03**
-  Россия возобновляет авиасообщение с Японией, Сербией и Кубой
РФ возобновит авиасообщение с Сербией, Кубой и Японией, а также увеличит количество рейсов в Швейцарию, Белоруссию, ОАЭ и на Мальдивы, сообщил оперативный штаб по борьбе с коронавирусом. ... Помимо этого, оперативный штаб принял решение увеличить число рейсов, совершаемых в такие страны, как Швейцария, Белоруссия, ОАЭ и Мальдивы. **Наша версия 16:02**
-  Россия возобновляет авиаперелеты с Японией, Кубой и Сербией
Оперативный штаб по борьбе с распространением коронавирусной инфекции сообщил о возобновлении Россией авиасообщения с Японией, Кубой и Сербией. ...Так, дважды в неделю из Москвы можно будет вылететь в Белград, Кайо-Коко и Санта-Клару (Куба). **Южный Федеральный 15:58**

Примеры задач

- рубрикация новостей

Главное Коронавирус Москва и область Интересное Политика Общество Экономика В мире Спорт

Компьютерная лингвистика. Приложения

Немного подробнее

- Язык:
 - две стороны
 - план выражения – план содержания



- Приложения:
 - автоматическая обработка плана выражения
 - автоматическая обработка «контента»

Компьютерная лингвистика. Приложения. “Мешок” задач

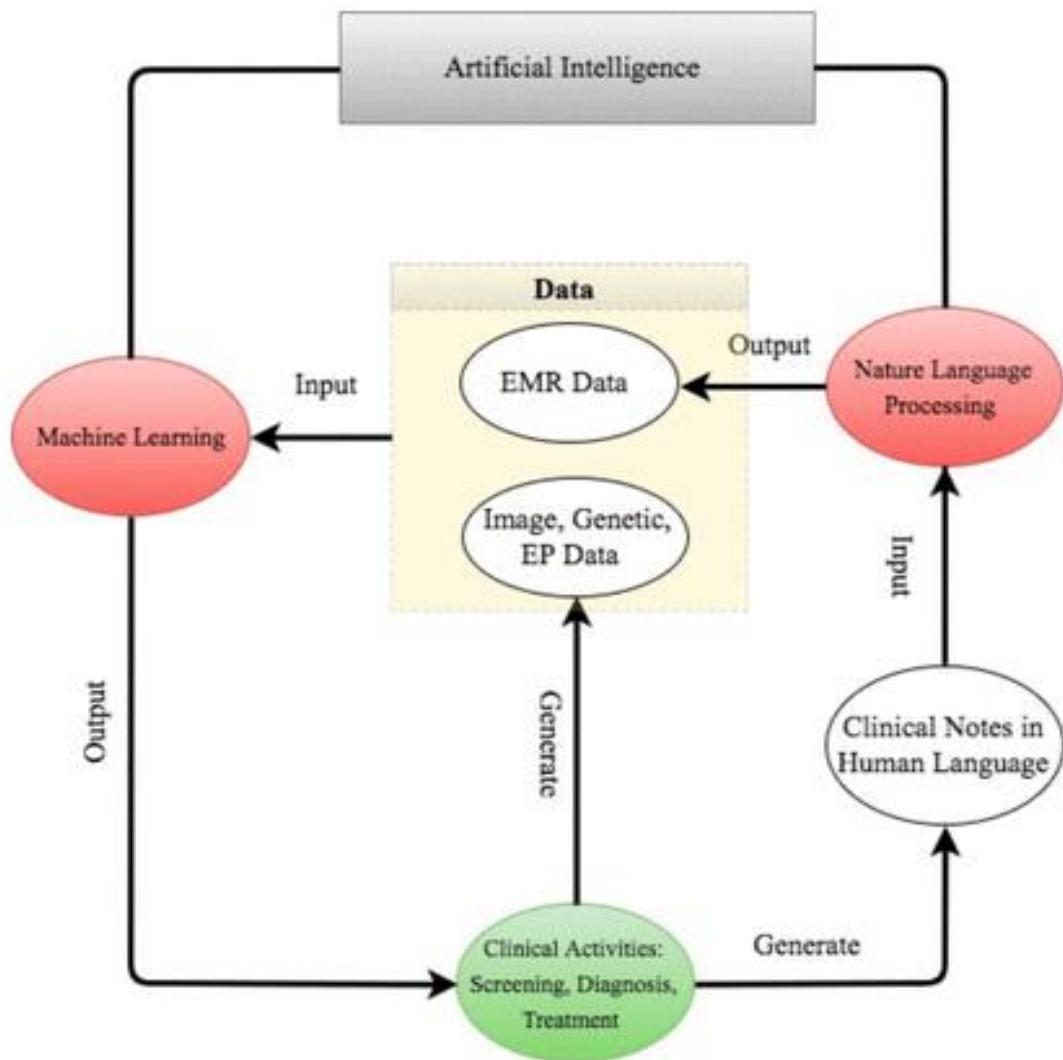
Занимается практическими задачами:

- Проверка правописания, грамматики и стиля.
- Распознавание текстов (печатный, рукописный).
- Распознавание (диктовка, слитная) и синтез речи.
- ---
- Машинный перевод текста и речи (классика NLP).
- Вопросно-ответные системы и системы логического вывода.
- Симплификация текста
- DH (Digital Humanities)
- Автоматические помощники (Алиса...)

Занимается практическими задачами:

- Поиск нужного документа по запросу (в т.ч. в Интернете).
- Реферирование (смысловое сжатие).
- Классификация (кластеризация) текстов по содержанию, установление сходства текстов (плагиат и т.п.).
- Автофилترация (определение нежелательных документов: спам и т.п.)
- Вопросно-ответные системы и системы логического вывода.
- Системы извлечения знаний (Text Mining, Information Retrieval), мнений (Opinion Mining, Sentiment Analysis)

Структурированные данные могут быть обработаны



Структурированные данные
встраиваются в другие процессы
(и это AI)

Задача NLP – из
неструктурированных массивов
данных извлечь
структурированные данные,
которые встраиваются в
различные задачи ИИ

Немного истории

Два направления как движущие силы развития автоматической обработки текста

(+ популяризации, а также важные точки в развитии Искусственного интеллекта как дисциплины)

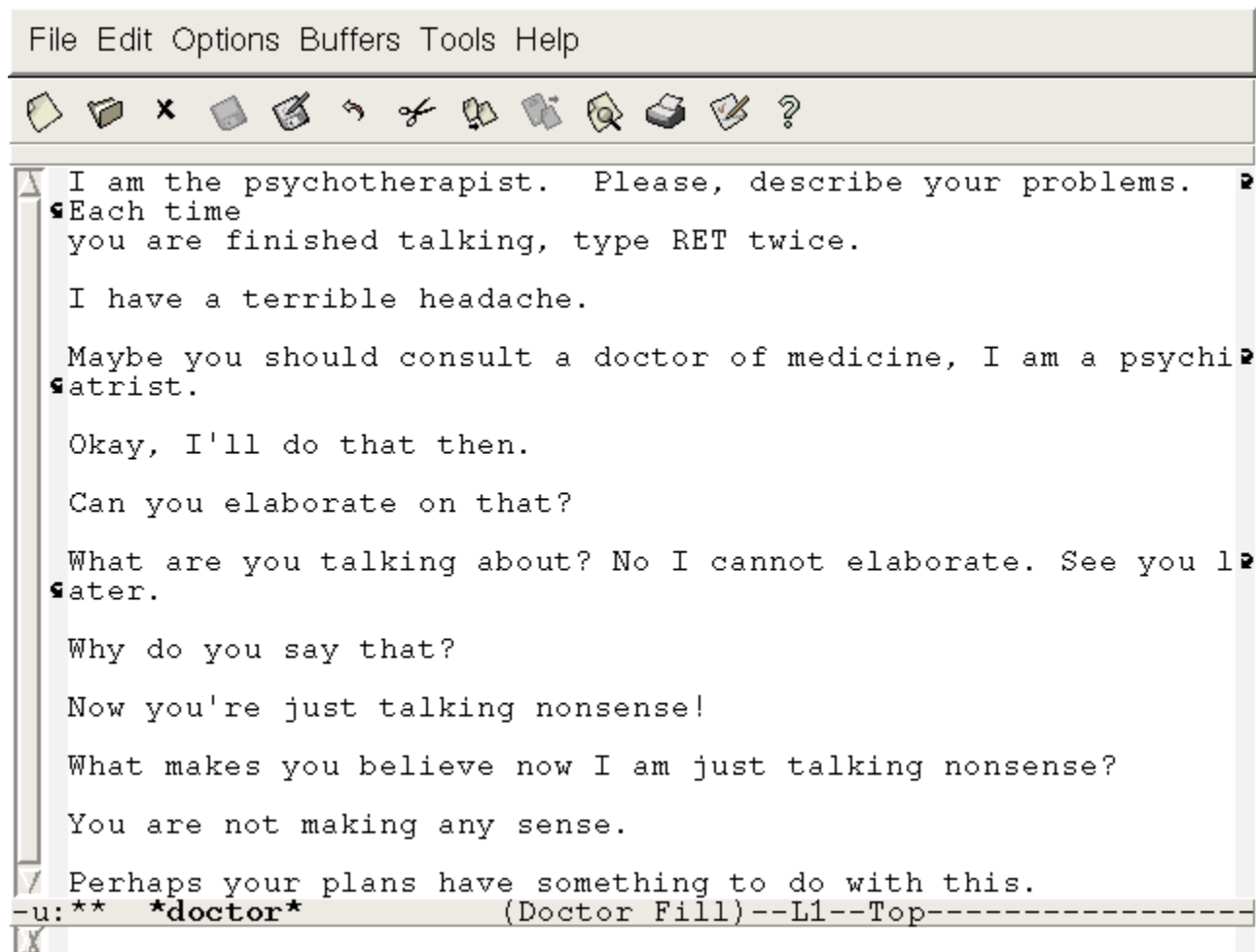
- ✓ машинный перевод
- ✓ диалоговые системы

Немного истории. Диалоговые системы

- **Тест Тьюринга**
- **1964-1966** – ELIZA, первые чатботы
- **1970е** – онтологии; Conceptual Dependency Theory (R. Schank)
- Развитие экспертных систем, систем, основанных на онтологическом моделировании ограниченной предметной области
- **конец 1980х-1990е** – внедрение статистических методов (распознавание речи, POS-tagging)

Немного истории. Диалоговые системы

- **1964-1966 – ELIZA,**
первые чатботы
 - Джозеф Вейценбаум
1966
 - Симулировала
интервью
психотерапевта с
пациентом
 - КЛЮЧЕВЫЕ СЛОВА + ШАБЛОНЫ



The screenshot shows a window titled "File Edit Options Buffers Tools Help" with a toolbar containing icons for file operations. The main text area displays a simulated conversation with a psychotherapist. The text is as follows:

```
I am the psychotherapist. Please, describe your problems.
Each time
you are finished talking, type RET twice.

I have a terrible headache.

Maybe you should consult a doctor of medicine, I am a psychiatrist.

Okay, I'll do that then.

Can you elaborate on that?

What are you talking about? No I cannot elaborate. See you later.

Why do you say that?

Now you're just talking nonsense!

What makes you believe now I am just talking nonsense?

You are not making any sense.

Perhaps your plans have something to do with this.
-u:** *doctor* (Doctor Fill)--L1--Top-----
```

Немного истории. Диалоговые системы

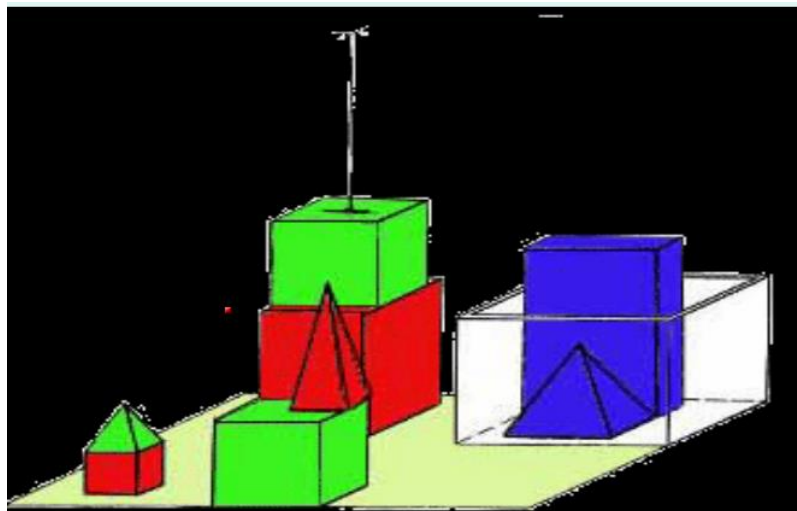
- Система Т. Винограда SHRDLU

Человек: *Возьми пирамиду.*

Компьютер: *Я не понимаю, какую пирамиду Вы имеете в виду.*

Человек (передумав): *Найди блок, который выше того, который ты держишь, и положи его в ящик.*

Компьютер: *Под «ним», полагаю, вы имели в виду блок, который выше того, что я держу.*



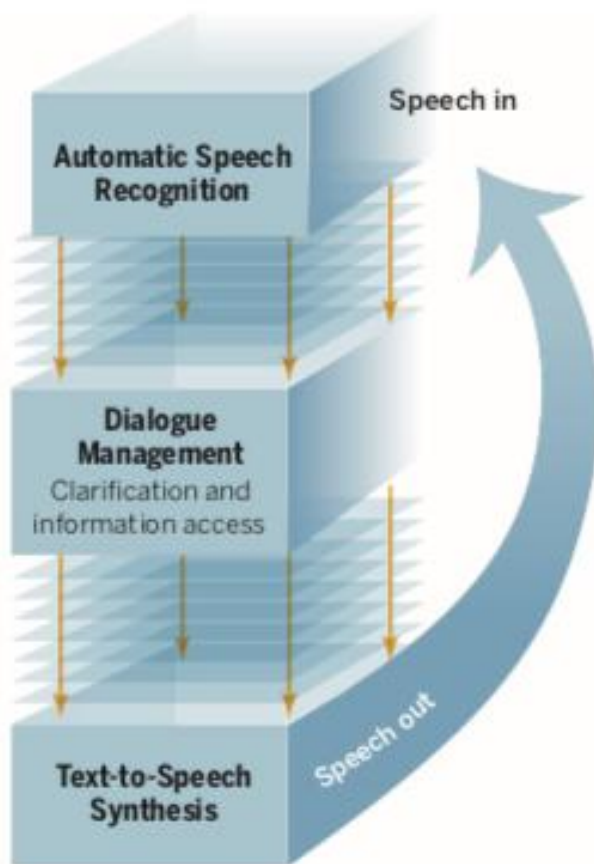
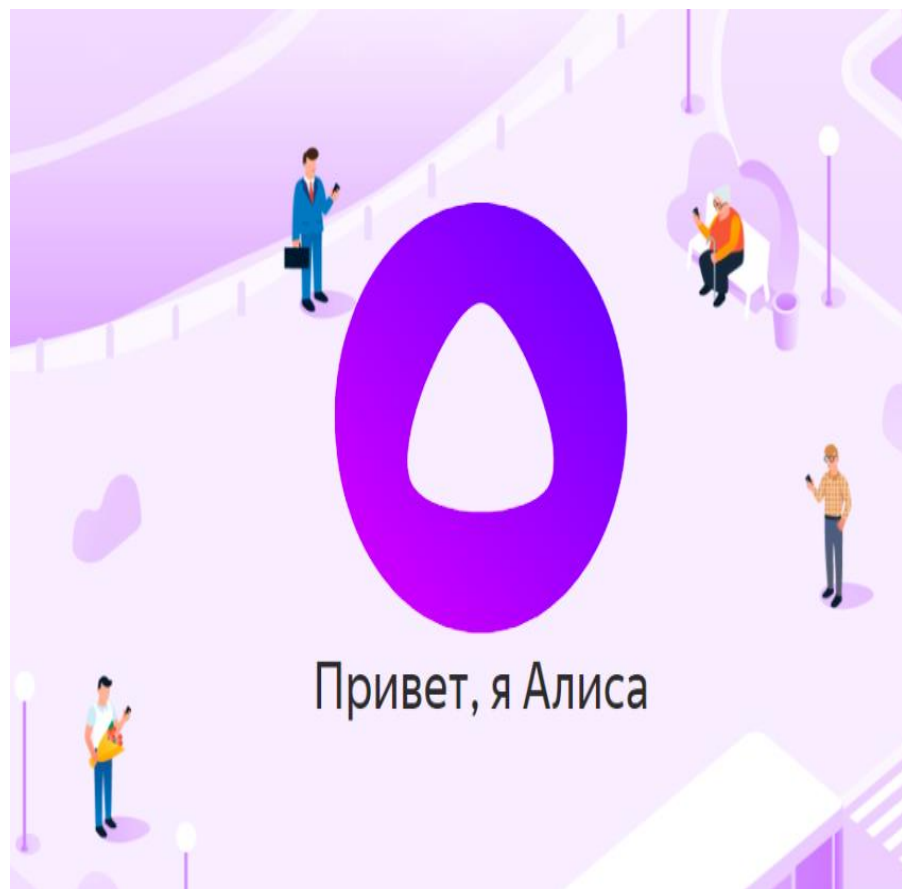


Fig. 3. A spoken dialogue system. The three main components are represented by rectangles; arrows denote the flow of information.

Hirschberg Manning 2015



Машинный перевод

- Джорджтаунский эксперимент: 7 января 1954 года в штаб-квартире IBM в Нью-Йорке (60 предложений с русского на английский)
- Машинный перевод на основе правил:
 - системы дословного перевода (Direct Machine Translation)
 - трансферные системы (Transfer-based Machine Translation: перевод целых фраз, учет идиом и т.п.)
 - интрелингвистические системы (перевод в некоторое концептуальное, независимое от конкретного языка представление), например, модель СМЫСЛ-ТЕКСТ – система ЭТАП
- статистический машинный перевод (Statistical machine translation): параллельные корпуса + выравнивание «фраз»
- нейронный машинный перевод (энкодер-декодер)

- Автоматический анализ текста
 - text mining
 - информационный поиск
 - извлечение знаний
 - общение с пользователем на естественном языке



- собственно лингвистическая обработка текста + решение задач извлечения информации из текста

- **Обработка коллекций текстов:**
 - Группировка текстов / разделение текстов / похожие тексты
- **Задачи анализа контента:**
 - Найти тексты, похожие по смыслу, стилю, тематике
- **Анализ текста:**
 - Извлечение/выделение фрагментов текста; извлечение онтологических элементов (элементов знаний); преобразование неструктурированных данных в структурированные
- **Задача анализа контента:**
 - Извлечение информации определенного типа из текста

- **Обработка коллекций текстов:**

- Группировка текстов / разделение текстов / похожие тексты

- **Задачи:**

- Найти тексты, похожие по смыслу, стилю, тематике
- Текст – объект обработки

- информационный поиск
- агрегация новостей
- рубрикация
- ...

Информационный поиск

[Washington area unemployment rate hovered at 5.4 percent in August](#)

Washington Post - 2 hours ago

The unemployment rate in the **Washington** area remained at 5.4 percent in August, even as jobs were added across a broad range of sectors, ...

[Unemployment falls in nearly 90 pct. of US cities](#) Bellingham Herald
[all 163 news articles »](#)

[Questions and answers on presidential debates](#)

Washington Post - 54 minutes ago

WASHINGTON — Tired of being deluged with TV commercials telling you that President Barack Obama or challenger Mitt Romney “approved ...

[Obama and Romney face off: Kathleen Parker, The Washington Post](#) NOLA.com

[National, Florida, Virginia poll numbers tighten; Obama up 8 in Ohio](#)

Washington Times (blog)

[Washington Informer](#)

Основная модель:

- Модель “информационного поиска”
- Основные допущения:
 - текст – объект – признаки - «мешок» слов (bag of words)
 - каждое слово появляется в тексте независимо от другого
- Текст – точка (вектор) в n-мерном пространстве
- Каждое измерение задается словом, которое есть в каком-нибудь тексте коллекции документов
- Близкие тексты имеют похожий набор слов
- Если текст – это вектор, то можно измерить расстояние между двумя текстами

- Основная модель:
 - модель “информационного поиска”
- Основные допущения:
 - текст – «мешок слов» (bag of words)
- каждое слово появляется в тексте независимо от другого -
 - > вероятность увидеть слово X в тексте не зависит от вероятности увидеть слово Y
- *Ворон к ворону летит, ворон ворону кричит*

Ворон к ворону летит, ворон ворону кричит

w_{11} w_{22} w_{13} w_{34} w_{15} w_{16} w_{57}

Объем текста: N – количество словоупотреблений в тексте –

Или $N = \sum_{i=1}^L fr(w_i) = 4+1+1+1 = 7$

L – количество разных (несовпадающих) слов в тексте (или объем словаря)

Вероятность увидеть словоупотребление *Ворон* (первое словоупотребление в тексте) $1/7$

$P(w_{ij}) = 1/7$ (i – номер слова в «словаре», j – порядковый номер словоупотребления в тексте)

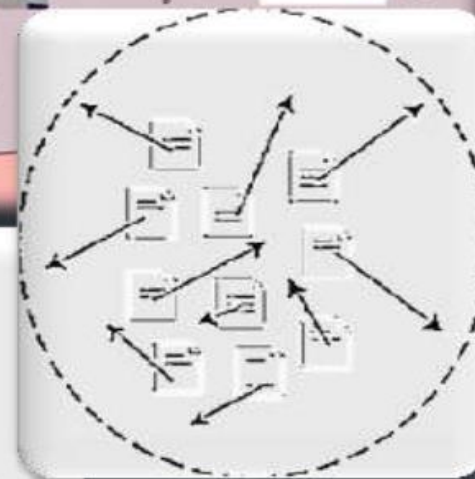
- *Ворон к ворону летит, ворон ворону кричит*
 - $w_{11} w_{22} w_{13} w_{34} w_{15} w_{16} w_{57}$
 - Вероятность увидеть слово *ворон* (лексему)
 - $P(w_{.i}) = Fr(w_{.i}) / N = 4/7$
-
- Дисклеймер к модели «мешок» слов:
 - событие ‘появление в тексте слово *лететь*’ не совсем независимо от события ‘появление в тексте слова *ворон*’

Обработка коллекция документов

- Содержание каждого документа или тематической коллекции документов (**семантический портрет**) может быть единообразно описано вектором семантически значимых компонентов в пространстве семантических признаков. Компоненты:
 - текстовые n-граммы,
 - объекты,
 - теги и др. метаданные.
- Вектора можно сравнивать по “похожести”, т.е. сравнивая контекст документов, можно вычислять расстояния и степень их близости, формируя кластеры близких документов.
- Сравнивая документы (кластеры документов) разнесенные во времени можно объединять их в сюжет и показывать ретроспективу его развития.

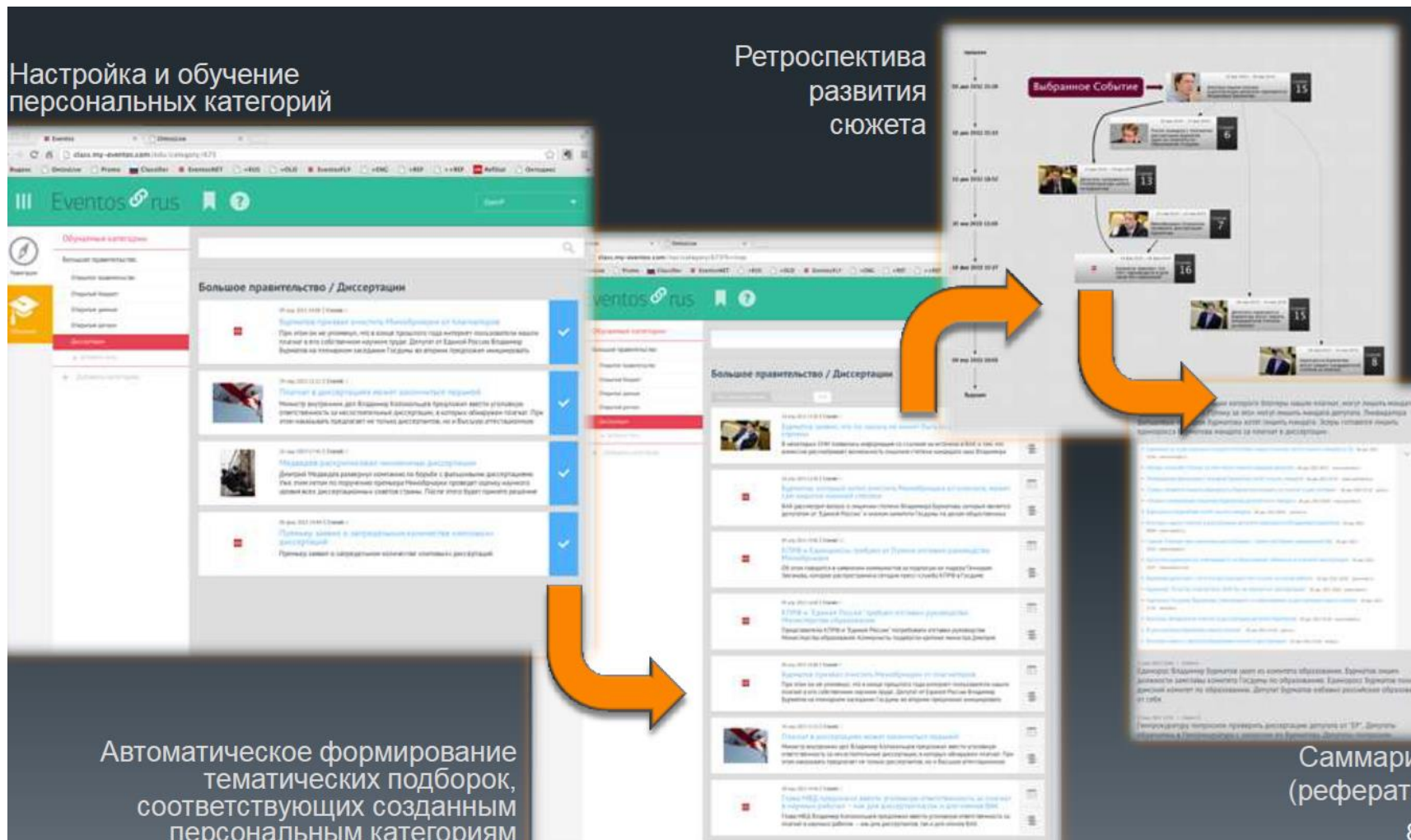


Косинусная мера близости векторов документов



Результат:
4 млн. документов
40 млн. словарь
400 тыс. сюжетов
75 тыс. ретроспектив
8 тыс. больших историй

Обработка коллекция документов





Основные задачи анализа контента

- **Анализ текста:**
 - Извлечение/выделение фрагментов текста; извлечение онтологических элементов (элементов знаний); преобразование неструктурированных данных в структурированные
- Задачи извлечения информации из текста (Information extraction):
 - Извлечение событий, их участников, места, времени, последовательности событий, отношений (Named Entities Recognition (Instances Extraction), relation extraction, fact extraction)
 - Извлечение онтологических знаний (knowledge extraction)

Извлечение информации из текста

- George Washington (February 22, 1732 – December 14, 1799), was one of the Founding Fathers of the United States, serving as the commander-in-chief of the Continental Army during the American Revolutionary War and later as the new republic's first President. He also presided over the convention that drafted the Constitution. Washington, D.C., the capital of the United States, is named for him, as is the State of Washington on the nation's Pacific Coast.*

Person /people/person		
edit	Date of birth:	Feb 22, 1732
edit	Place of birth:	<div>location</div> <div>Colonial Beach</div> <div>contained by</div> <div>Westmoreland County</div> <div>Virginia</div> <div>United States of America</div>
edit	Country of nationality:	Kingdom of Great Britain, United States of America
edit	Gender:	Male
edit	Profession:	Surveyor, Engineering, Politician, Farmer, Soldier, Military officer
edit	Religion:	Episcopal Church in the United States of America, Anglicanism, Deism

Извлечение информации из текста



- **George Washington** (*February 22, 1732 – December 14, 1799*), was one of the Founding Fathers of the **United States**, serving as **the commander-in-chief** of the Continental Army during the American Revolutionary War and later as the new republic's first President. He also presided over the convention that drafted the Constitution. Washington, D.C., the capital of the United States, is named for him, as is the State of Washington on the nation's Pacific Coast.



Person /people/person			
edit	Date of birth:	Feb 22, 1732	
edit	Place of birth:	location	contained by
		Colonial Beach	Westmoreland County Virginia United States of America
edit	Country of nationality:	Kingdom of Great Britain, United States of America	
edit	Gender:	Male	
edit	Profession:	Surveyor, Engineering, Politician, Farmer, Soldier, Military officer	
edit	Religion:	Episcopal Church in the United States of America, Anglicanism, Deism	

<http://www.freebase.com/>

Извлечение информации из текста

Objects	Properties	Text
<p>All</p> <p>Person</p> <ul style="list-style-type: none"> Обама Барак Патрушев Николай ✓ Путин Владимир Шойгу Сергей Джордж Литтл Кэйтлин Хэйдэн Томас Донилов Юрий Ушаков <p>Географическое ...</p> <ul style="list-style-type: none"> Вашингтон Ирландия Москва Оклахома Россия США 	<p>Патрушев Николай</p> <p>imageUrl </p> <p>givenName Николай</p> <p>additionalName Платонович</p> <p>lastName Патрушев</p> <p>статья Википедии http://ru.wikipedia.org/wiki/Николай_Патрушев</p> <p>birthDate 11.07.1951</p> <p>birthPlace Ленинград</p> <p>EmployedBy Совет Безопасности РФ</p> <p>EmployedBy ФСБ РФ</p> <p>title Герой Российской Федерации</p> <p>title генерал армии</p> <p>title доктор юридических наук</p>	<p>Политика </p> <p>Развитие отношений между США и Россией, в том числе в экономической области, обсудили президент Барак Обама и секретарь Совета безопасности РФ Николай Патрушев. Встреча состоялась 22 мая в Белом доме и на ней, помимо прочего, затронули вопросы борьбы с терроризмом и ситуацию в Сирии. Как сообщила официальный представитель Совета национальной безопасности США Кэйтлин Хэйдэн, Обама заглянул на встречу Патрушева с помощником президента США Томасом Дониловым. Президент США подтвердил желание укреплять двусторонние отношения, в том числе американо-российские экономические связи. Они также говорили о важности углубления сотрудничества в борьбе с терроризмом и необходимости политического урегулирования в Сирии путем переговоров",</p>

Извлечение информации из текста: онтология DBPedia

Тексты бывают разные

About: George Washington

An Entity of Type : [Concept](#), from Named Graph : <http://dbpedia.org>, within Data Space : [dbpedia.org](#)

Property	Value
rdf:type	▪ skos:Concept
rdfs:label	▪ George Washington
owl:sameAs	▪ http://pl.dbpedia.org/resource/Kategoria:George_Washington ▪ http://de.dbpedia.org/resource/Kategorie:George_Washington ▪ http://fr.dbpedia.org/resource/Catégorie:George_Washington ▪ http://ko.dbpedia.org/resource/분류:조지_워싱턴 ▪ http://ru.dbpedia.org/resource/Категория:Джордж_Вашингтон
skos:broader	▪ category:Presidents_of_the_United_States ▪ category:Washington_family ▪ category:Wikipedia_categories_named_after_American_politicians
skos:prefLabel	▪ George Washington
http://www.w3.org/ns/prov#wasDerivedFrom	▪ http://en.wikipedia.org/wiki/Category:George_Washington?oldid=490387017
is dcterms:subject of	▪ dbpedia:Washington's_Birthday ▪ dbpedia:United_States_presidential_election_1788-1789 ▪ dbpedia:1932_Washington_Bicentennial ▪ dbpedia:Conway_Cabal ▪ dbpedia:Potomac_Company ▪ dbpedia:List_of_places_named_for_George_Washington

Тексты бывают разные

- Подписи под фотографиями:



<http://www.flickr.com/photos/tags/animals/>

✓ Как должны быть устроены ярлыки, чтобы можно было найти фотографии по одной теме?

Актуальные направления: анализ социальных сетей

- большие данные, огромная заложенная ценность
- новый, нестандартный языковой материал (code switching, hate speech, языковая креативность)
- приватность личных данных - палка о двух концах

- Распознавание состояния участников коммуникации (Speakers states)
- анализ тональности
- степень уверенности

часто объединяется с анализом социальных медиа

Актуальные направления

Были сегодня первый раз !

Все очень понравилось .

И место на диванчиках нашлось , и официантка очень приветливая , но ненавязчивая , интерьер приятный , живая музыка тоже расслабляет , вполне и под нее можно вести бес

Обслуживание оперативное , ждать 1-2 минуты пиво , гренки чуть дольше , пиво отлич

!

!

я в восторге !

!

!

обстановка очень уютная .

а еще радует что недалеко от дома такое прекрасное заведение есть .

category

Food positive

Interior positive

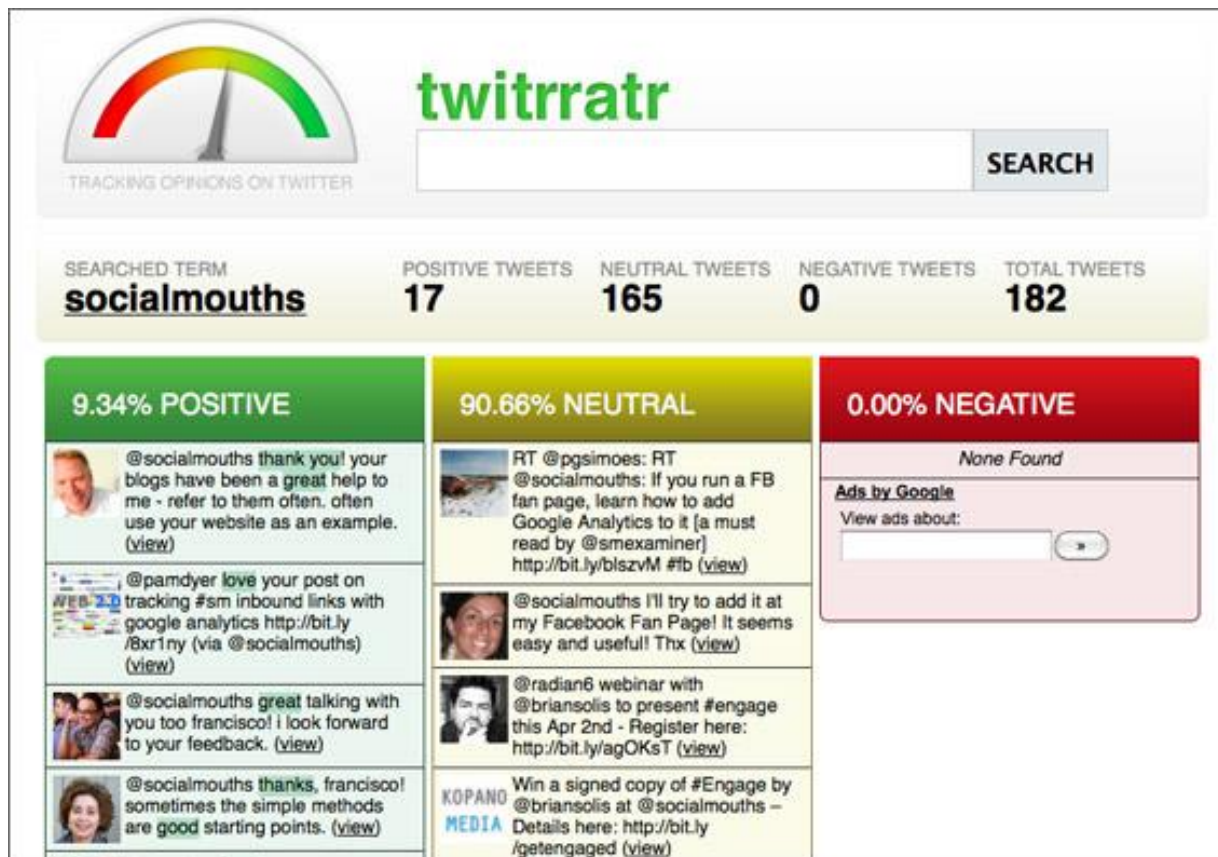
Price absence

Service positive

Whole positive

- Судя *по официальным данным*, которые огласил начальник отдела Минфина Петр Казакевич, *по состоянию* на 24 июля заработано более 85 млрд руб (3,4 млрд долларов), а доходность в долларах составила 9,5 процента. *Эксперты уже наперебой говорят* о блестящих результатах при инвестировании в самые надежные бумаги
- Субъективная информация

Извлечение тональности (сентимент). Twitter



<http://socialmouths.com/2010/03/31/6-tools-for-twitter-sentiment-tracking/>

ВЫЯВЛЕНИЕ ИНФОРМАЦИОННЫХ ТРЕНДОВ

ПРОЕКТ «ОЛИМПИЙСКОЕ НАСТРОЕНИЕ». СОВМЕСТНО С РИА НОВОСТИ

Мониторинг по видам спорта на русском и английском языках.

Выявляемые тренды:

- **Популярные персоны и виды спорта:** определяем среди лидеров роста в рейтинге популярных слов;
- **Пики количества обсуждений;**
- **Тональность обсуждений.**

Ежедневный поток 1-2,5 млн сообщений.



- **Обработка коллекций текстов:**
 - Группировка текстов / разделение текстов
- Задачи анализа контента:
 - Найти тексты, похожие по смыслу, стилю, тематике
- Задачи обработки текстов:
 - Найти тексты, похожие на некоторый текст (например, запрос пользователя) – информационный поиск;
 - Собрать похожие тексты в одну группу - новостная агрегация, удаление дублей – кластеризация текстов
 - «рассортировать» тексты по группам – рубрикация текстов, классификация по стилям, распознавание спама

- **Анализ текста:**

- Извлечение/выделение фрагментов текста; извлечение онтологических элементов (элементов знаний); преобразование неструктурированных данных в структурированные

- **Задача анализа контента:**

- Извлечение событий, их участников, места, времени, последовательности событий, отношений
- Извлечение оценки событий, объектов, мнений
- Семантическое аннотирование
- Извлечение онтологических знаний

(Named Entities Recognition (Instances Extraction), Fact Extraction, Relation Extraction, Semantic Annotation, Ontological Information Extraction)

- **Резюме**
- **Три направления:**
 - **Ресурсы и инструменты для изучения языка**
 - корпуса, лексикографические ресурсы (словари, тезаурусы), специальные программы для работы с корпусами, обработки звука
 - **Формальные модели**
 - Двухуровневая морфология (конечные преобразователи); формальный синтаксис (грамматики и т.п.)
 - **Автоматическая обработка текста**
 - обработка коллекций текстов: информационный поиск, рубрикация, новостная и др. агрегация;
 - извлечение информации: извлечение именованных сущностей, извлечение фактов, извлечение мнений;
 - др.

Введение в компьютерную лингвистику

Немного терминологии

Корпуса	Corpora
Аннотирование корпуса	Corpus annotation (tagging)
Автоматическая обработка ЕЯ (текста) / АОТ	Natural language processing (NLP)
Информационный поиск	Information retrieval (IR)
Извлечение информации из текста	Information extraction (IE)
Извлечение именованных сущностей	Named entities extraction (NER)
Извлечение фактов	Fact extraction
Тематическая группировка текстов (анализ новостного потока)	Topic detection and tracking (TDT)
Автоматическая рубрикация текстов	Text classification
Модель мешка слов	Bag of words