

# Natural language processing

## 1. Intro

Katya Chernyak

CS HSE

September 6, 2018

# Overview

## Why take this course?

Natural Language Processing (NLP) is one of the most interesting and research intensive fields of modern Artificial Intelligence. Obviously NLP has a lot of industry and business applications (try to name a few!).

# What this course is about?

## ① Basics

- ▶ Vector space model and dimensionality reduction
- ▶ Word embeddings
- ▶ Probabilistic language models
- ▶ Probabilistic sequence models: HMM, CRF
- ▶ Neural sequence models: RNN, biLSTM
- ▶ Text classification: fasttext, CNN
- ▶ Named entity recognition: CNN-biLSTM-CRF

## ② Advanced topics

- ▶ seq2seq models, GAN, VAE, Transformer
- ▶ Universal models: ELMo, ULMFiT

## ③ Industry

- ▶ Machine translation
- ▶ Speech recognition
- ▶ Spelling correction
- ▶ Conversational intelligence

## ④ Research

- ▶ Compositional semantics
- ▶ Graph embeddings

# Course information

- ① Repo: github
- ② Tlg: (?)
- ③ Final mark:  $0.4 * \text{HW} + 0.2 * \text{essay} + 0.4 * \text{exam}$

# Lecture schedule (1 module)

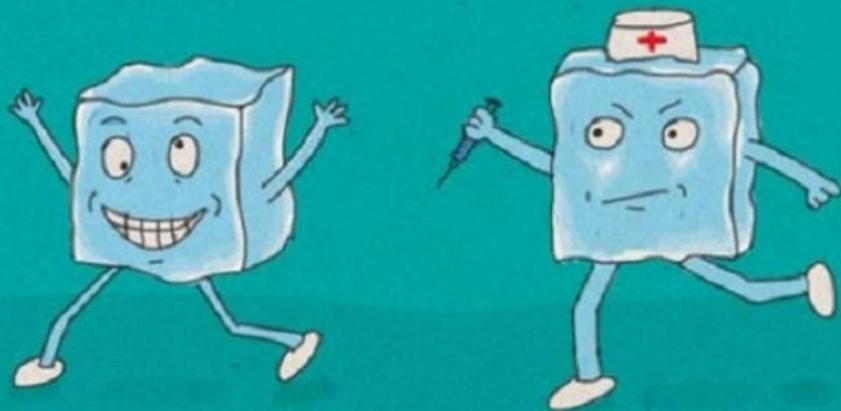
- ① Week 1 Intro + Regexp
- ② Week 2 VSM + key words and phrases extraction
- ③ Week 3 Practical. Russian POS taggers and text classification
- ④ Week 4 VSM and dimensionality reduction + topic modelling
- ⑤ Week 5 Distributional semantics + w2v
- ⑥ Week 6 Language models + text generation

# Overview

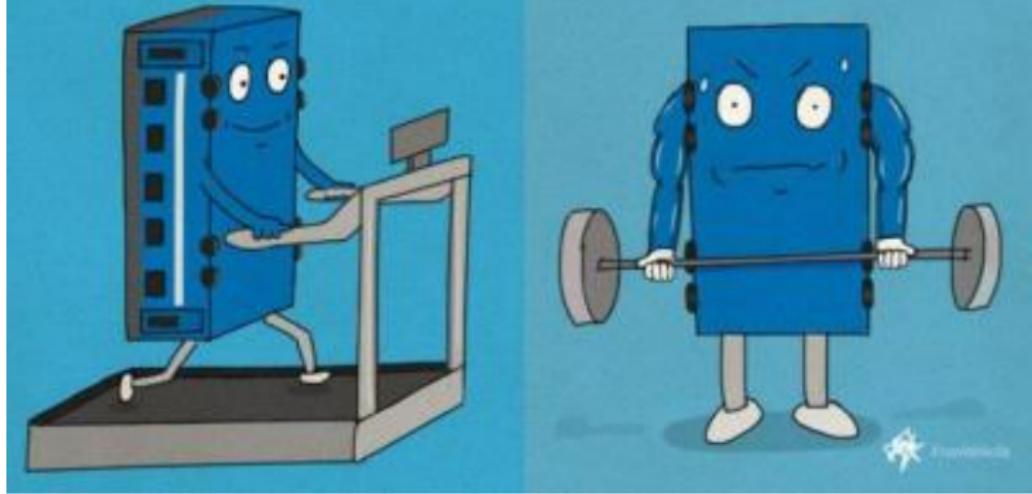
СТЕПАН ПЕТРОВИЧ  
ПРОЦВЕТАЕТ



Лёд, тронулся...

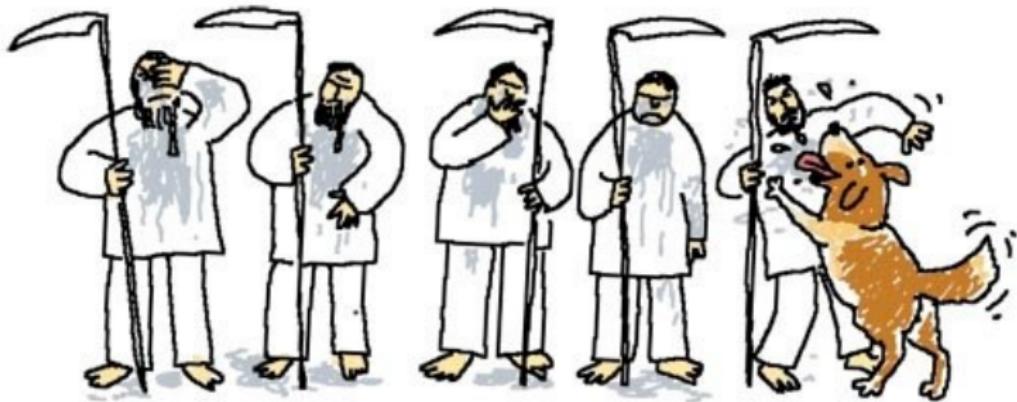


Голубой вагон  
бежит — качается

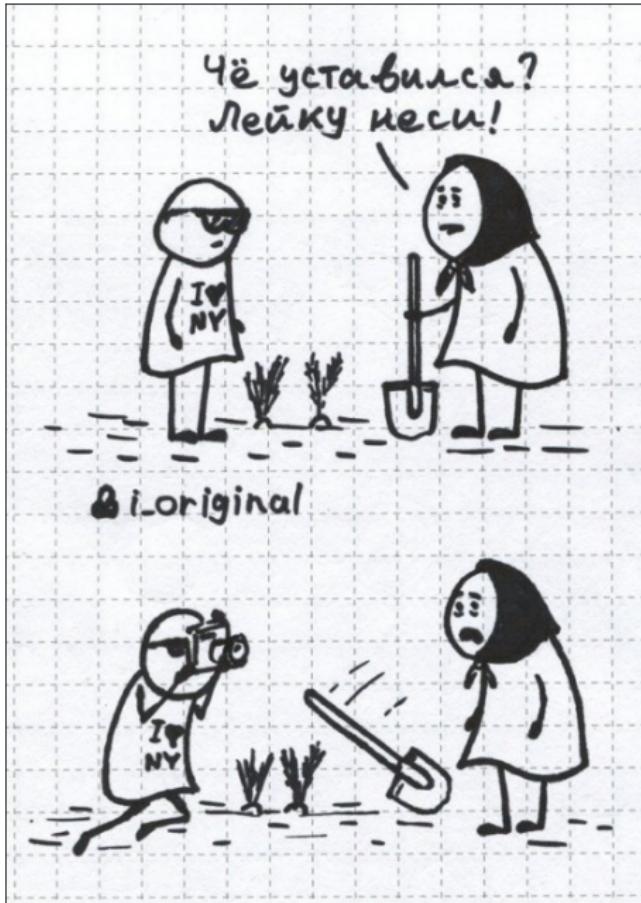




ОБЛАКА КЛУБЯТСЯ



Отслюнявил пять косарей



МАЛЬВИНА ГЕНРИХОВНА ЗАЩИЩАЕТ ДОКТОРСКУЮ





Messages

Lyss

Edit

Call

Contact Info

Sep 29, 2011 10:24 AM

I'm so glad you taught me  
how to orgasm. 😊

Omg omg omg omg.  
What?!?!

Lol

Organize! I meant  
organize!! Haha.

Oh my phone!! Haha

Haha I needed that laugh  
today. Thank you  
iPhones!!



DAMNYOUAUTOCORRECT.COM

Send

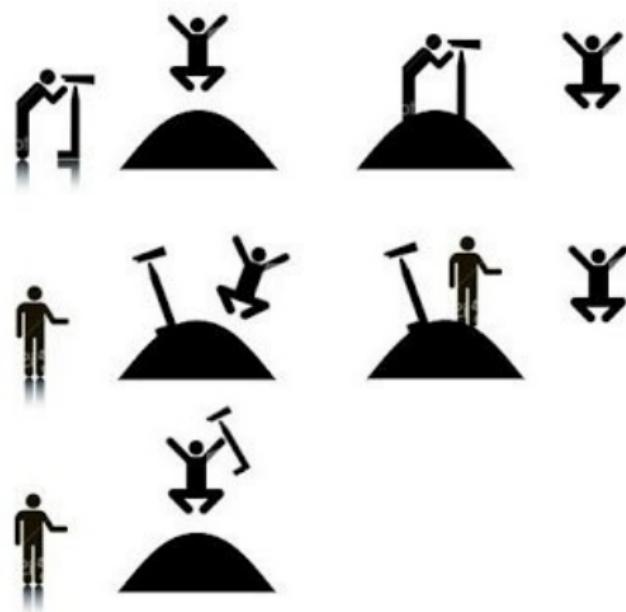


# Levels of analysis

- ① Tokenization
- ② Lemmatization and POS-tagging
- ③ Parsing

# Ambiguity

- ➊ Polysemy and word-sense disambiguation: орган, bank
- ➋ Homonymy: the ship or to ship, стекло
- ➌ Syntactic ambiguity: John saw the man on the mountain with a telescope.



# Multilingual NLP<sup>1</sup>

A

Наташа купила водку  
Natasha purchased the Vodka.

B

Наташу купила водка  
The vodka purchased Natasha.



RusVectores nearest neighbors:

Google nearest neighbors:

<sup>1</sup><https://primer.ai/blog/Russian-Natural-Language-Processing/>

# Multilingual NLP<sup>2</sup>

водки	vodka	водкой	vodka
водка»	vodka»	газировка	soda-water
водку	vodka	безалкогольная	nonalcoholic
водка –	vodka –	водку»	vodka»
водке	vodka	водки»	vodka»

Facebook nearest neighbors:

газировка	soda-water	лимонада	lemonade
безалкогольная	non-alcoholic	квас	kvass
слабоалкогольный	low-alcohol	рюмка	shot-glass
бутылка	bottle	пивобезалкогольный	non-alcoholic beer
напитки	drinks	десертная	dessert

Facebook-lem nearest neighbors:

# Overview

# Language resources

- ① Corpus: BNC, HKРЯ
- ② WordNet, RuWordNet
- ③ TreeBanks, Universal Dependencies
- ④ Task-specific annotated collections for classification, question answering, named entity recognition, NL inference, image captioning, TempEval, SenseEval, etc
- ⑤ Social media
- ⑥ Web as Corpus projects
- ⑦ Wikipedia

## Результаты поиска в основном корпусе

[перейти на страницу поиска](#) [выбрать подкорпус](#) [версия с ударениями](#) [настройки](#) [обычный формат](#) [English](#)

Объем всего корпуса: 115 645 документов, 23 803 881 предложение, 283 431 966 слов.

## истребитель

Найдено 4 056 вхождений.

[Распределение по годам](#) [Статистика](#)Поискать в других корпусах: [акцентологическом](#), [газетном](#), [драматическом](#), [мультиязычном](#), [обучающем](#), [параллельном](#), [поэтическом](#), [синтаксическом](#), [устном](#).Страницы: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [11](#) [следующая страница](#)

парка истребительной авиации РФ составляют **истребители** широком диапазоне высот полёта, управлять **истребителями** пущенными ЗУР осуществляется с борта **истребителя** разведки и информационными бортовыми средствами **истребителей** источниками информации могут оказаться БРЛС **истребителей** полностью выдачи информации от БРЛС **истребителей** и дальности выдачи информации от **истребителей** поддержка со стороны АКРДДН и **истребителей** проблем управления ЗУР с борта **истребителей** Проблема управления ЗУР с борта **истребителей** в прошлом профессионально лётавший на **истребителях** самолётов фронтовой авиации, снайд— самолётов, **истребителем** и впёрёд за неизвестным богом— **истребителем** но их могли перехватить ночные **истребители** свою очередь, помогут нам модернизировать **истребитель** назвал создание радара "Жук" для **истребителя** В частности, для модернизации **истребителя** и все 18 пусковых установок. **Истребитель** дальнего действия МиГ-31, **истребители** и обмениваться с ними боевой (ИОС-III). достаточно высок и практически может и АКРДДН. на КП ЗРС "Триумф", на могут быть сокращены в том может заключаться в выдаче информации —перехватчиков. требует решения принципиально новых задач обладает большим потенциалом и амбициями и истребителей-бомбардировщиков. спотыкаясь на разбитые скворища — вместе или встретить зенитным огнём над МиГ-29. F 8- ПМ по заказу МИГ-29 создан радар "Жук" спутников В 60-е в СССР

## Результаты поиска в основном корпусе

[перейти на страницу поиска](#) [выбрать подкорпус](#) [версия с ударениями](#) [настройки](#) [обычный формат](#) [English](#)

Объем всего корпуса: 115 645 документов, 23 803 881 предложение, 283 431 966 слов.

**истребитель**

Найдено 4 056 вхождений.

[Распределение по годам](#) [Статистика](#)
Поискать в других корпусах: [акцентологическом](#), [газетном](#), [диалектном](#), [мультимедийном](#), [обучающем](#), [параллельном](#), [поэтическом](#), [синтаксическом](#), [устном](#).Страницы: [препылаша страница](#) [83](#) [84](#) [85](#) [86](#) [87](#) [88](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [следующая страница](#)

все гости. — Честь и слава **истребителю** предрассудков! — проревел один толстый англичанин и пожар, как два ангела-**истребителя** , протекали Дюовек из конца в природы сильнее законов приличия. — Лучший **истребитель** время! — прымолила Марина. видел тоже и знаменитого Гуссейна, **истребителя** янчар. а в военное, как ангел-**истребитель**, являлся с своими крылатыми полками Как ангел-**истребитель** летел перед своим отрядом Юрий ), — не поэт, а **истребитель** бумаги и усыпил? на шаха; но сей безжалостный **истребитель** единокровных умел явить себя великим ревностно служив обманщику, венанция его **истребителей** и желая очаровать их, писали ангел-**истребитель** посетил кладовые, смотрению твоему вверенные

## Текстовый вход: ИСТРЕБИТЕЛЬ

### САМОЛЕТ-ИСТРЕБИТЕЛЬ

(ИСТРЕБИТЕЛЬ, ИСТРЕБИТЕЛЬНЫЙ КОМПЛЕКС, САМОЛЕТ-ИСТРЕБИТЕЛЬ)

ВЫШЕ БОЕВОЙ САМОЛЕТ

ЦЕЛОЕ ИСТРЕБИТЕЛЬНАЯ АВИАЦИЯ

АССОЦ<sub>2</sub> БАРРАЖИРОВАНИЕ

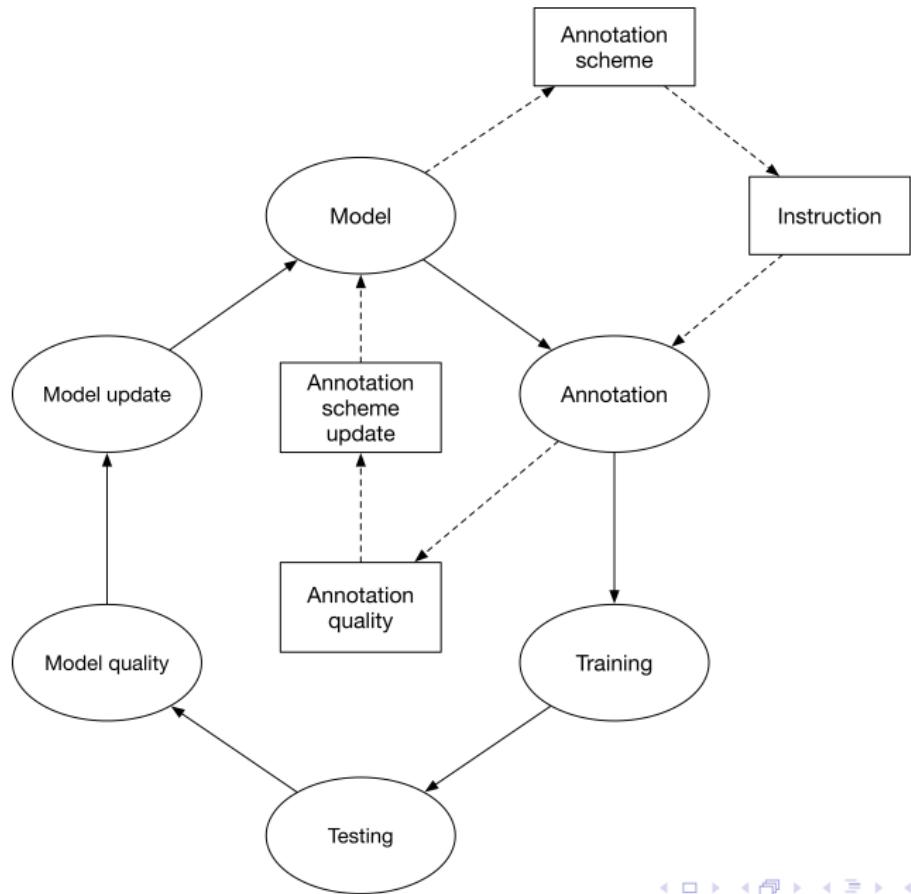
### ИСТРЕБИТЕЛЬ (ТОТ, КТО ИСТРЕБЛЯЕТ)

(ИСТРЕБИТЕЛЬ)

ВЫШЕ ЖИВОЙ ОРГАНИЗМ

ЦЕЛОЕ ИСТРЕБИТЬ

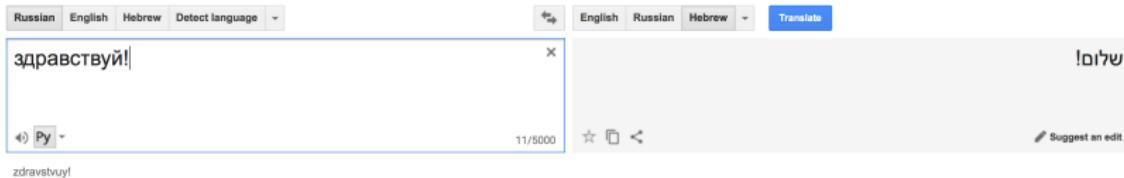
# NLP Pipeline



# Overview

# Industry

## Machine translation



## Text classification

- ▶ Spam filtering
- ▶ By topic
- ▶ By sentiment

A screenshot of a Gmail inbox titled 'All Mail'. The sidebar shows categories: Spam (176), Trash, and Categories (Social: 10, Promotions: 16, Updates: 189, Forums). The main area displays three spam emails from 'Blade Runner 2049 (2017)' with subject lines 'Бегущий по лезвию 2049 (2017)', 'Терраин башни (2017)', and 'Донжон (2017)'. Each email has a green 'Mark as spam' button below it. To the right of the emails are icons for 'Like', 'Dislike', and '5 октября 2017' (October 5, 2017). The bottom of the screen shows the standard Gmail navigation bar with icons for back, forward, search, and refresh.

- Text clustering

Сейчас в СМИ в Москве 5 сентября, среда 09:20

- Г На Урале произошло землетрясение магнитудой 5,4
- Опубликовано видео удара Израиля по Сирии
- СМИ сообщили о появлении дырки в «Союзе» при сборке в Королеве
- Т Лавров заявил, что конфронтация между РФ и США нарастает
- В России предложили установить минимальную цену на пиво

USD MOEX 68,11 **-0,01** EUR MOEX 78,91 **-0,23** НЕФТЬ 77,79 **-0,33 %** ...

# Industry

- Information extraction

- Named entity recognition
- Facts and events



Yaroslav Kuzminov Share

**Born:** May 26, 1957 (age 60), [Moscow](#)

**Spouse:** [Elvira Nabillina](#)

**Children:** [Ivan Kuzminov](#), [Angelina Yaroslav](#), [Vasily Kuzminov](#)

**Institution:** [National Research University Higher School of Economics](#)

Profiles

 Twitter

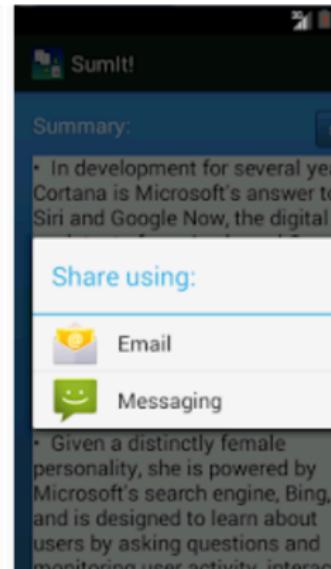
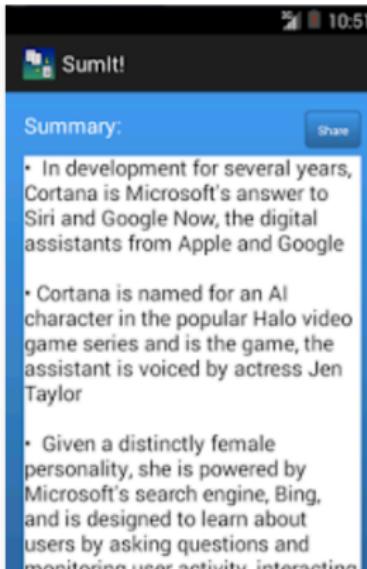
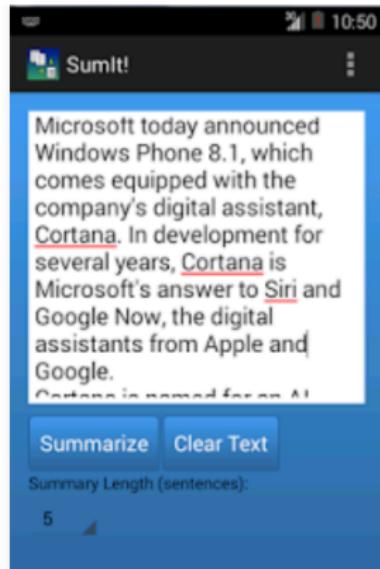
# Industry

- Chat-bots
- Question-answering systems
- Cognitive assistants

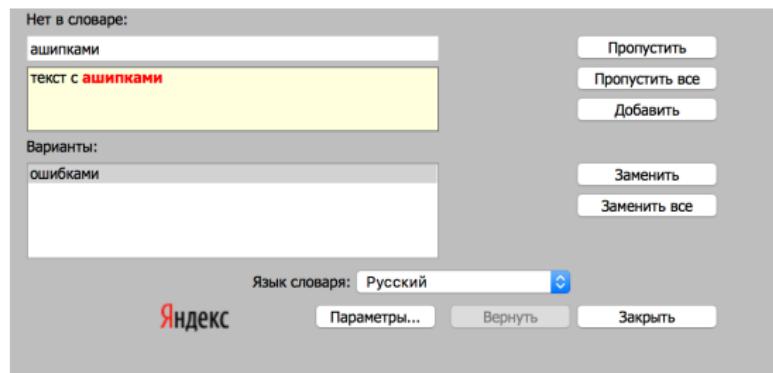


# Industry

- Text summarization

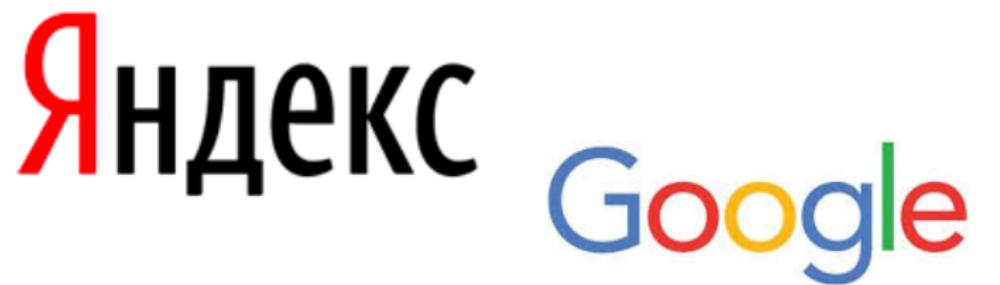


- Speech recognition (speech to text, STT) and speech synthesis (text to speech, TTS)
- Optical character recognition (OCR)
- Spell checking



## Industry

- Information retrieval and web search engines



# Overview

# ML tasks

## ① Unsupervised tasks

- ▶ Word models
- ▶ Sentence models
- ▶ Topic models

## ② Supervised tasks

- ▶ Text classification
- ▶ Sequence labelling and structured prediction
- ▶ Sequence-to-sequence models
- ▶ Tree models
- ▶ Advanced learning techniques: RL, GAN, VAE, Bayesian methods

# Tools

- Scraping: API, scrapy, lxml
- Preprocessing: NLTK, SpaCy , regexp
- Rule-based parsers: Tomita-parser, Yargs
- VSM: gensim
- Morphology: mystem3, pymorphy2
- Syntax: UDPipe, SyntaxNet
- Your favorite ML and DL frameworks
- Pretrained models (incl. rusvectores.org)

# Reading

- ① Stuart Russell, Peter Norvig. Artificial Intelligence: A Modern Approach, Ch. 21, Ch. 23
- ② Amber Stubbs, James Pustejovsky. Natural Language Annotation for Machine Learning