MIPT
MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY

АТП кафедра Алгоритмов и Технологий Программирования

SBER DEVICES

Intro to NLP

MIPT
4.02.2021
Anton Emelianov, Alena Fenogenova.

2021

https://github.com/king-menin/mipt-nlp2021

Today

- Intro
- About the course
- Recent trends in NLP
- Practice: Basic text processing

Course instructors

Anton Emelianov, login-const@mail.ru, @king_menin

* Ph.D. Candidate and Lecturer at Algorithms and Programming Technologies dept, MIPT,
* R&D NLP, SberDevices, Sber.
* RussianNLP team member.
* Linkedin: https://goo.gl/M5UtBr

Alena Fenogenova, alenush93@gmail.com

* Master degree in Computational linguistics at the Higher School of Economics
* R&D NLP, SberDevices, Sber.
* RussianNLP team member.
* Linkedin https://www.linkedin.com/in/alena-fenogenova-3b447417a/

Course information

- Github repo: https://github.com/king-menin/mipt-nlp2021
- Tlg chat: https://t.me/joinchat/HlYsCUgkZ9sTL0mJ
- Final mark: *score(HWs)/count(HWs)*10* or exam
- Homeworks: ~4, anytask - coming soon

Natural language processing

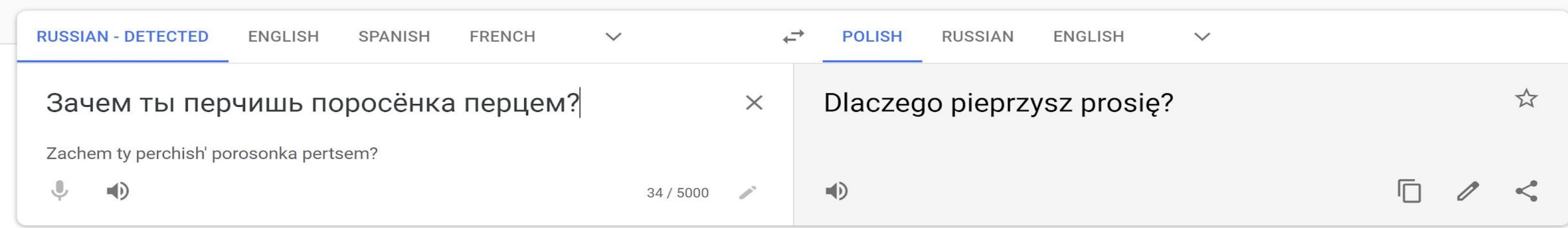
- **Natural Language Processing** (**NLP**) is a one of the most interesting and research intensive fields of modern Artificial Intelligence. Obviously NLP has a lot of industry and business applications.

Course Structure

Num	Title	Practical
1	Intro to NLP	Basic text processing
2	Word embeddings	Word2vec, text classification
3	RNN, CNN, text classification	LSTM, CNN
4	Language modeling, Part 1	Named entity recognition (NER)
5	Seq2Seq and machine translation	Seq2seq
6	Language modeling, Part 2 - attention, transformers	
7	Transfer Learning in NLP	Transformers model for classification
8	Syntax parsing	Syntax parsing
9	Question Answering	
10	Summarization and Simplification	
11	Real cases and trends	
12	TODO	

Industry

- **Machine translation**

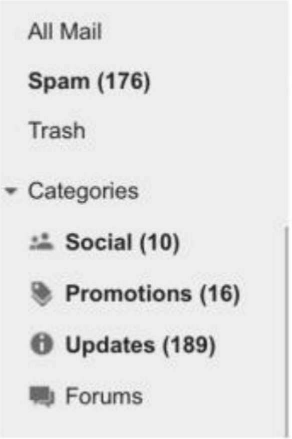


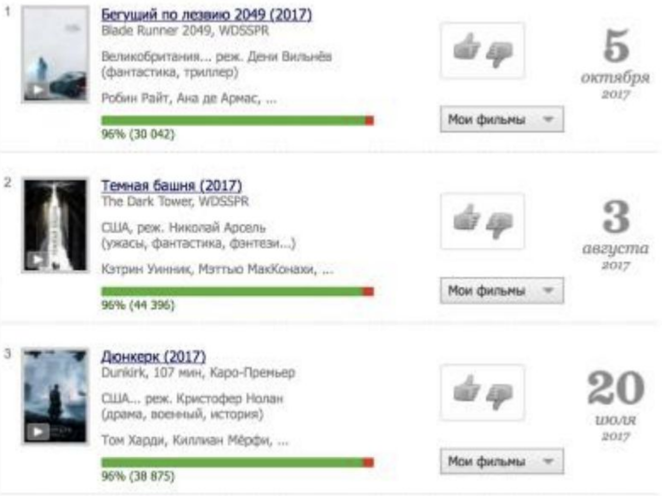
Зачем ты перчишь поросёнка перцем?

Zachem ty perchish' porosonka pertsem?

Dlaczego pieprzysz prosię?

- **Classification**
  - Spam filtering
  - Sentiment
  - Topic or genre





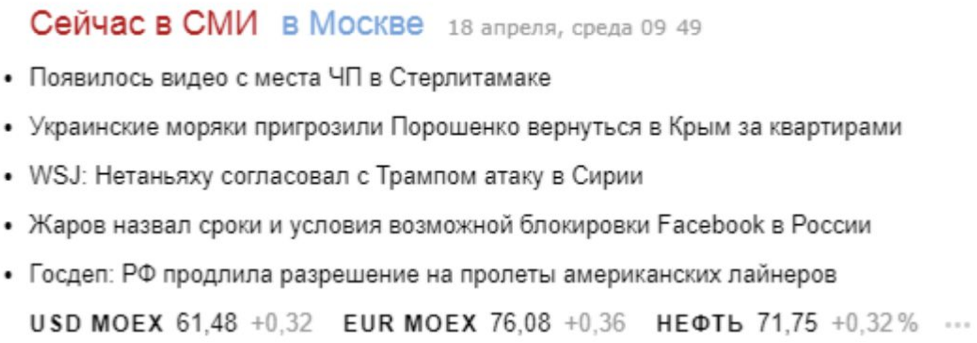
1 Бегущий по лезвию 2049 (2017)
Blade Runner 2049, WDSSPR
Великобритания... реж. Дени Вильнёв
(фантастика, триллер)
Робин Райт, Ана де Армас, ...
96% (30 042)
5 октября 2017

2 Темная башня (2017)
The Dark Tower, WDSSPR
США, реж. Николай Арсель
(ужасы, фантастика, фэнтези...)
Кэтрин Уинник, Мэттью МакКонахи, ...
96% (44 396)
3 августа 2017

3 Дюнкерк (2017)
Dunkirk, 107 мин, Каро-Премьер
США... реж. Кристофер Нолан
(драма, военный, история)
Том Харди, Киллиан Мёрфи, ...
96% (38 875)
20 июля 2017

Industry

- **Clustering texts**

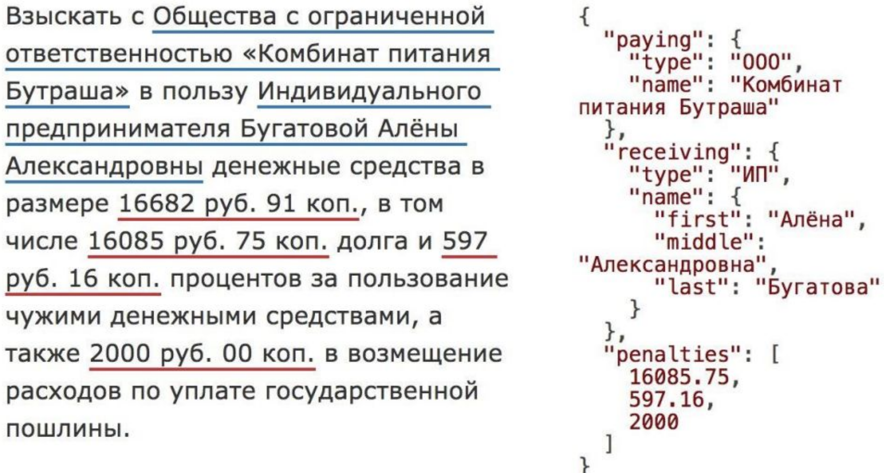


Сейчас в СМИ в Москве 18 апреля, среда 09 49

- Появилось видео с места ЧП в Стерлитамаке
- Украинские моряки пригрозили Порошенко вернуться в Крым за квартирами
- WSJ: Нетаньяху согласовал с Трампом атаку в Сирии
- Жаров назвал сроки и условия возможной блокировки Facebook в России
- Госдеп: РФ продлила разрешение на пролеты американских лайнеров

USD MOEX 61,48 +0,32 EUR MOEX 76,08 +0,36 НЕФТЬ 71,75 +0,32 % ···

- **Information extraction**
  - Facts
  - Events
  - NER

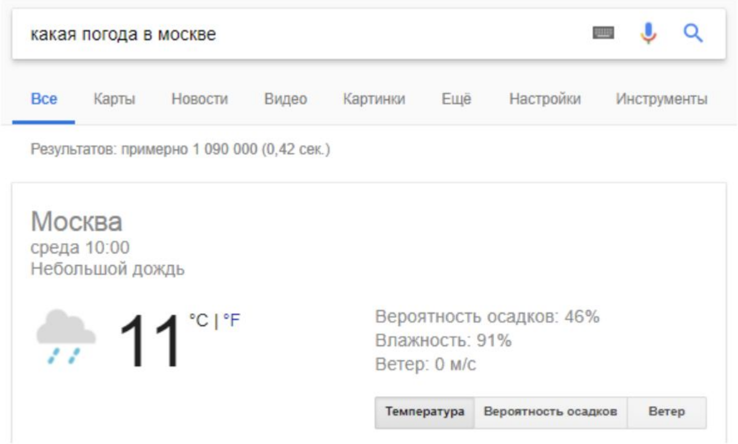


Взыскать с Общества с ограниченной ответственностью «Комбинат питания Бутраша» в пользу Индивидуального предпринимателя Бугатовой Алёны Александровны денежные средства в размере 16682 руб. 91 коп., в том числе 16085 руб. 75 коп. долга и 597 руб. 16 коп. процентов за пользование чужими денежными средствами, а также 2000 руб. 00 коп. в возмещение расходов по уплате государственной пошлины.

```
{
  "paying": {
    "type": "ООО",
    "name": "Комбинат
питания Бутраша"
  },
  "receiving": {
    "type": "ИП",
    "name": {
      "first": "Алёна",
      "middle":
"Александровна",
      "last": "Бугатова"
    }
  },
  "penalties": [
    16085.75,
    597.16,
    2000
  ]
}
```

Industry

- **Question Answering systems**



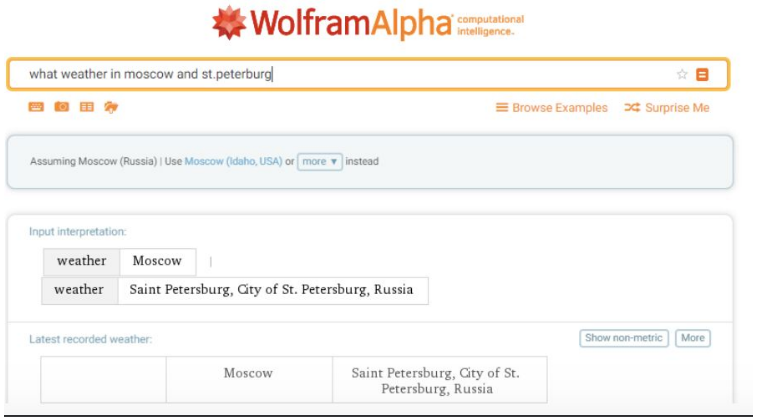
какая погода в москве

Результатов: примерно 1 090 000 (0,42 сек.)

Москва
среда 10:00
Небольшой дождь

11 °C | °F

Вероятность осадков: 46%
Влажность: 91%
Ветер: 0 м/с



what weather in moscow and st.peterburg

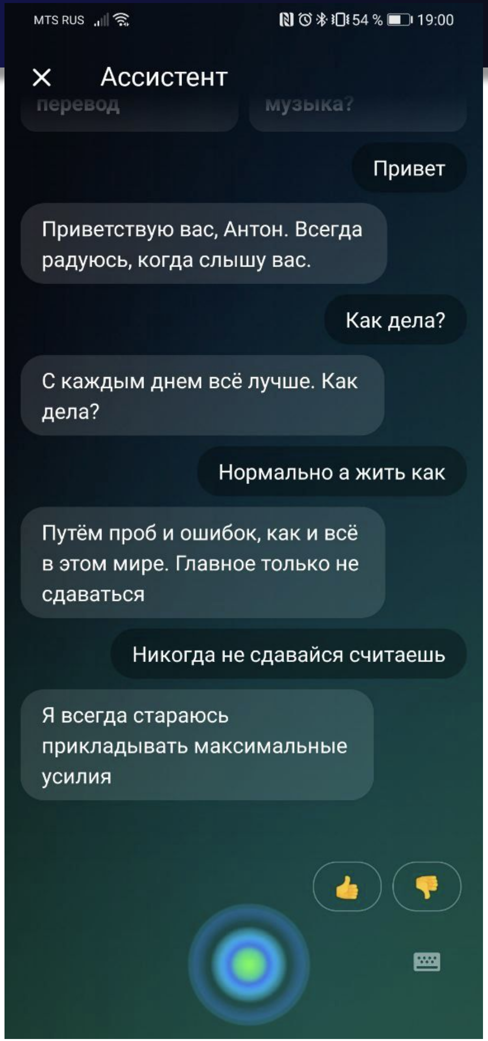
Assuming Moscow (Russia) | Use Moscow (Idaho, USA) or more ▾ instead

Input interpretation:

weather	Moscow
weather	Saint Petersburg, City of St. Petersburg, Russia

Latest recorded weather:

	Moscow	Saint Petersburg, City of St. Petersburg, Russia



Привет

Приветствую вас, Антон. Всегда радуюсь, когда слышу вас.

Как дела?

С каждым днем всё лучше. Как дела?

Нормально а жить как

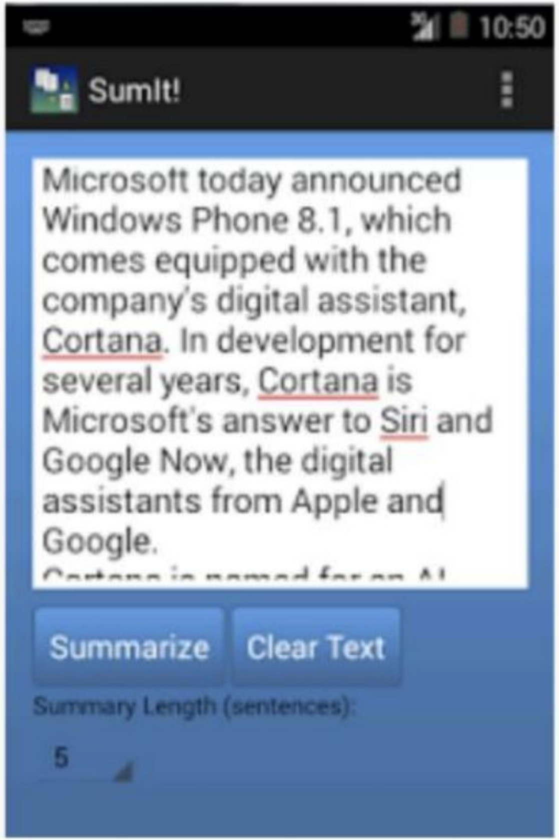
Путём проб и ошибок, как и всё в этом мире. Главное только не сдаваться

Никогда не сдавайся считаешь

Я всегда стараюсь прикладывать максимальные усилия

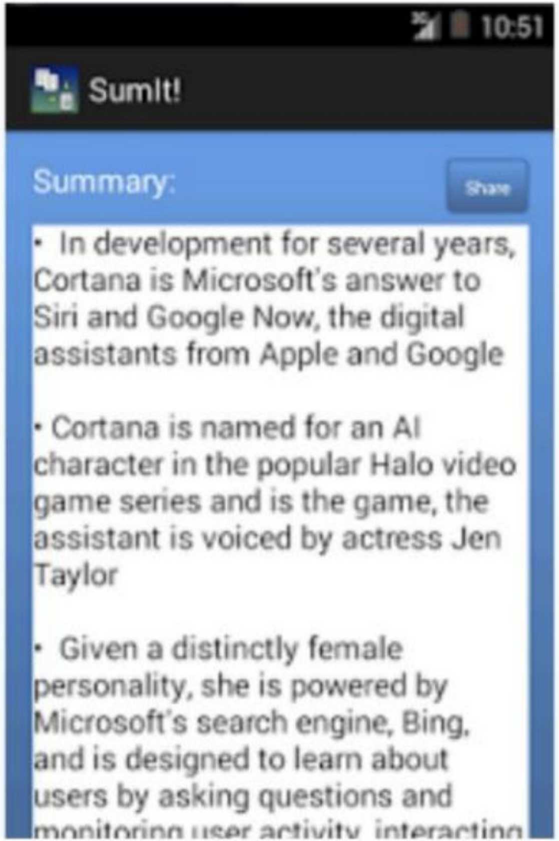
Industry

- **Text summarization**



Microsoft today announced Windows Phone 8.1, which comes equipped with the company's digital assistant, Cortana. In development for several years, Cortana is Microsoft's answer to Siri and Google Now, the digital assistants from Apple and Google.
Cortana is named for an AI

Summary Length (sentences):
5

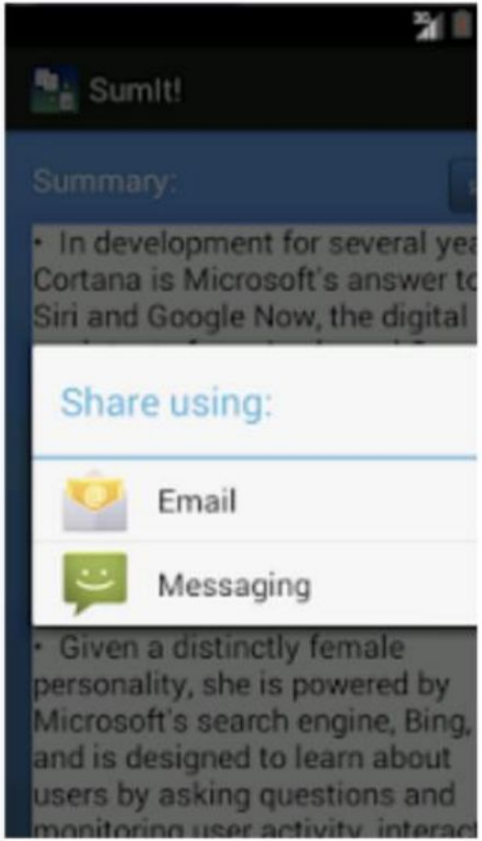


Summary:

- In development for several years, Cortana is Microsoft's answer to Siri and Google Now, the digital assistants from Apple and Google

- Cortana is named for an AI character in the popular Halo video game series and is the game, the assistant is voiced by actress Jen Taylor

- Given a distinctly female personality, she is powered by Microsoft's search engine, Bing, and is designed to learn about users by asking questions and monitoring user activity, interacting



Summary:

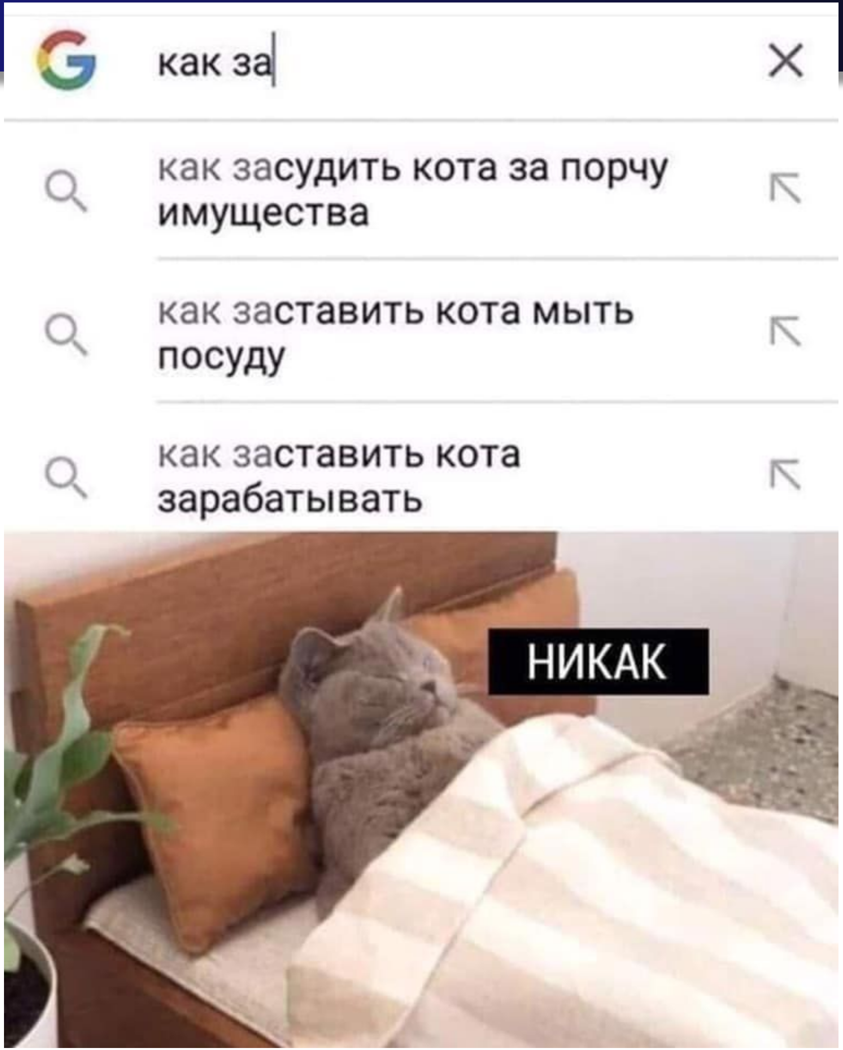
- In development for several yea
Cortana is Microsoft's answer to
Siri and Google Now, the digital

Share using:
Email
Messaging

- Given a distinctly female personality, she is powered by Microsoft's search engine, Bing, and is designed to learn about users by asking questions and monitoring user activity, interact

Industry

- **Text autocomplete**



как за

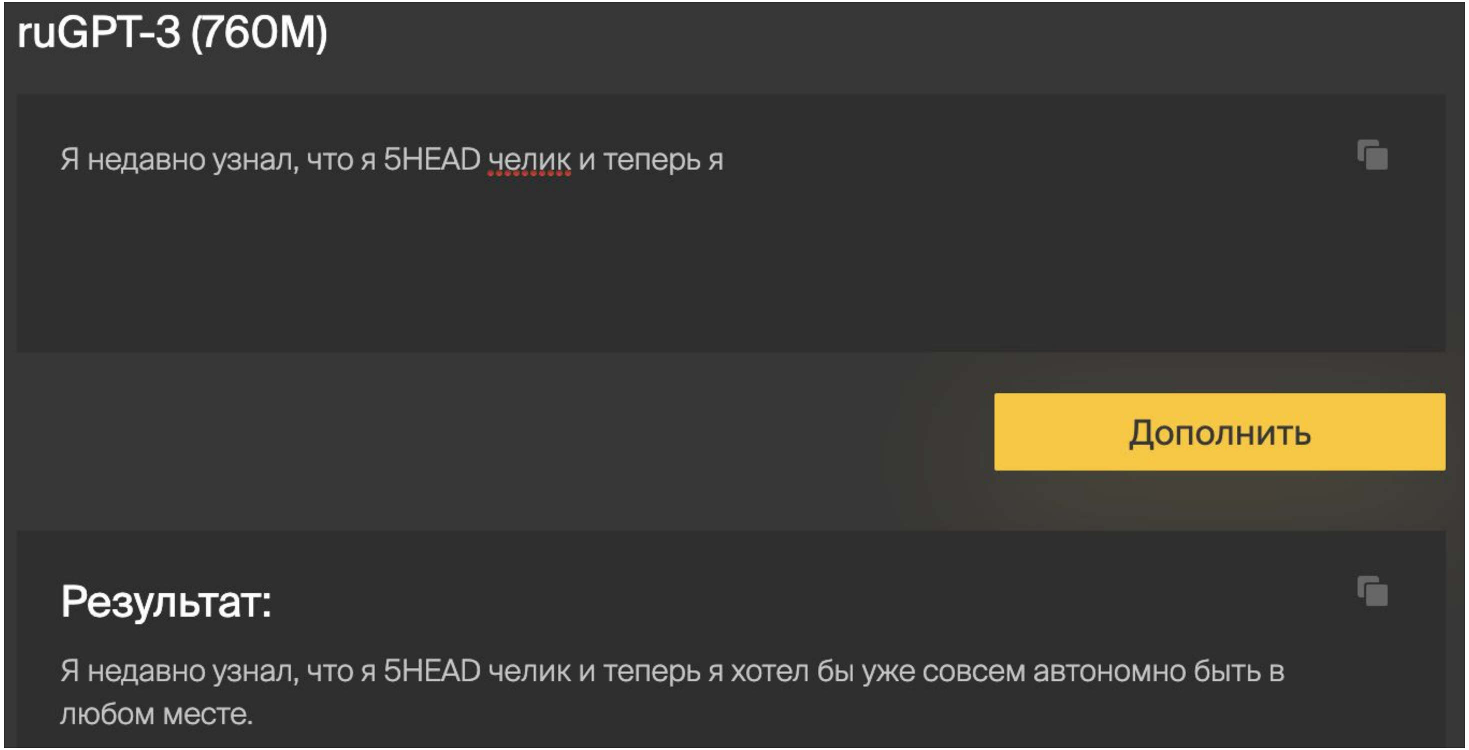
как засудить кота за порчу имущества

как заставить кота мыть посуду

как заставить кота зарабатывать

Industry

- **Text generation**



ruGPT-3 (760M)

Я недавно узнал, что я 5HEAD челик и теперь я

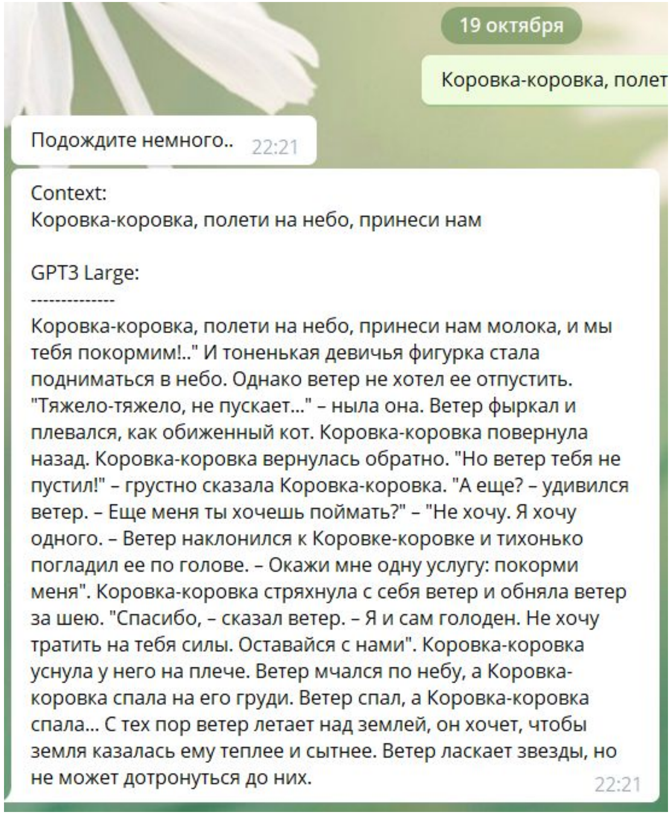
Результат:

Я недавно узнал, что я 5HEAD челик и теперь я хотел бы уже совсем автономно быть в любом месте.

https://sbercloud.ru/ru/warp/gpt-3

Industry

- **Text generation**



19 октября

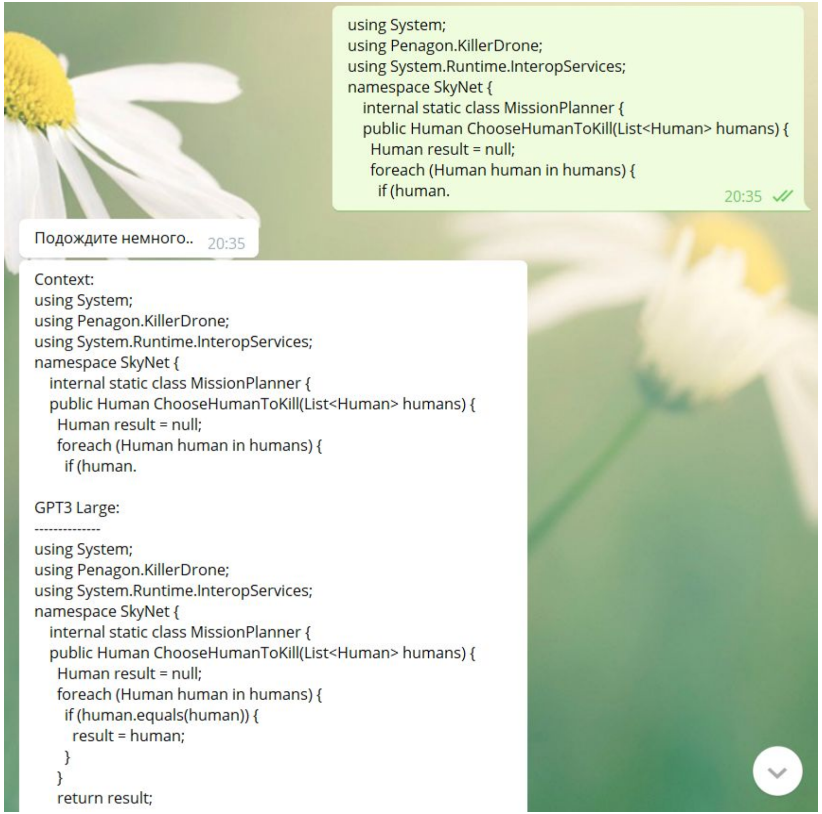
Коровка-коровка, полет

Подождите немного.. 22:21

Context:
Коровка-коровка, полети на небо, принеси нам

GPT3 Large:

Коровка-коровка, полети на небо, принеси нам молока, и мы тебя покормим!.." И тоненькая девичья фигурка стала подниматься в небо. Однако ветер не хотел ее отпустить. "Тяжело-тяжело, не пускает..." – ныла она. Ветер фыркал и плевался, как обиженный кот. Коровка-коровка повернула назад. Коровка-коровка вернулась обратно. "Но ветер тебя не пустил!" – грустно сказала Коровка-коровка. "А еще? – удивился ветер. – Еще меня ты хочешь поймать?" – "Не хочу. Я хочу одного. – Ветер наклонился к Коровке-коровке и тихонько погладил ее по голове. – Окажи мне одну услугу: покорми меня". Коровка-коровка стряхнула с себя ветер и обняла ветер за шею. "Спасибо, – сказал ветер. – Я и сам голоден. Не хочу тратить на тебя силы. Оставайся с нами". Коровка-коровка уснула у него на плече. Ветер мчался по небу, а Коровка-коровка спала на его груди. Ветер спал, а Коровка-коровка спала... С тех пор ветер летает над землей, он хочет, чтобы земля казалась ему теплее и сытнее. Ветер ласкает звезды, но не может дотронуться до них. 22:21



```
using System;
using Penagon.KillerDrone;
using System.Runtime.InteropServices;
namespace SkyNet {
   internal static class MissionPlanner {
   public Human ChooseHumanToKill(List<Human> humans) {
    Human result = null;
    foreach (Human human in humans) {
     if (human.
```

20:35 ✓✓

Подождите немного.. 20:35

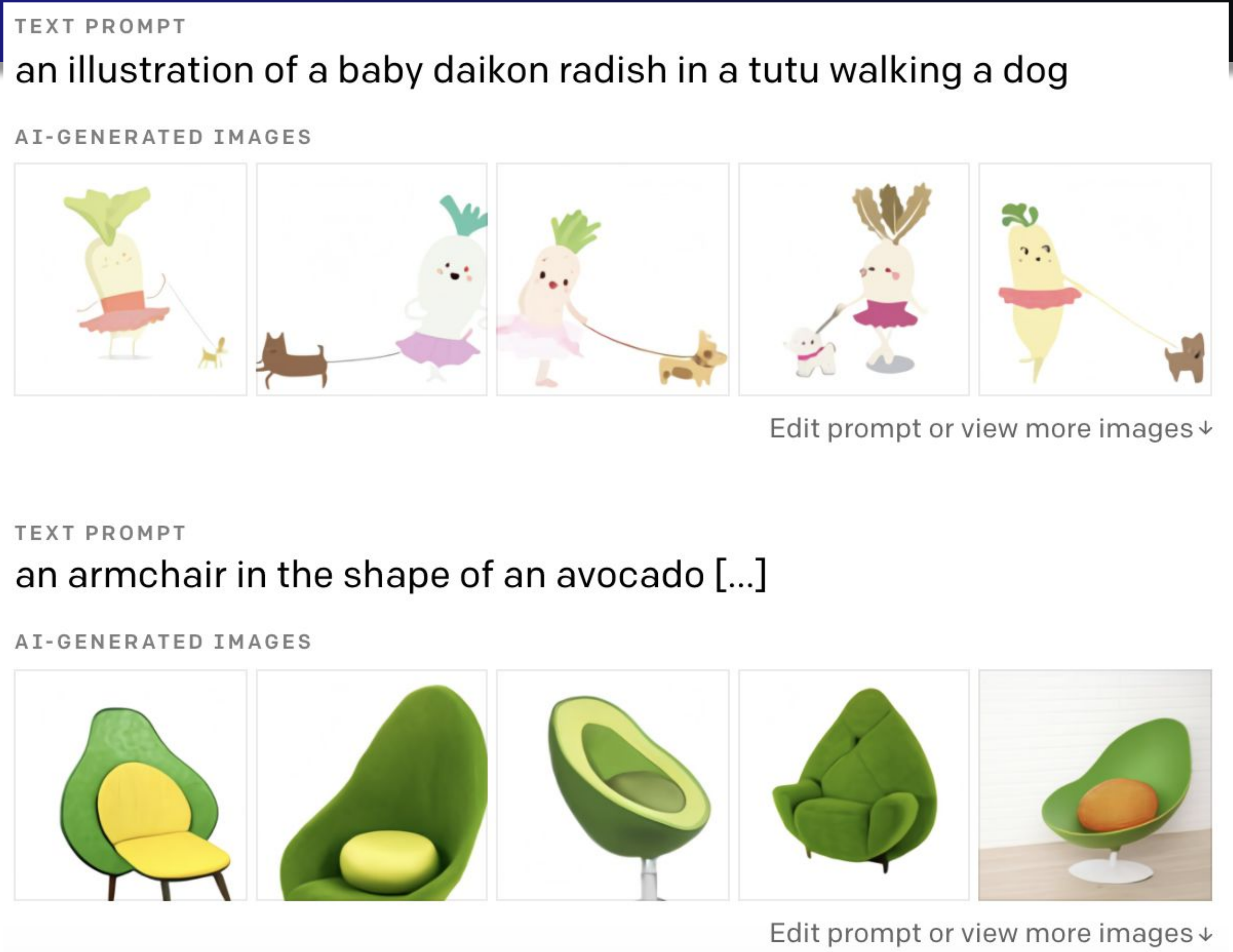
```
Context:
using System;
using Penagon.KillerDrone;
using System.Runtime.InteropServices;
namespace SkyNet {
   internal static class MissionPlanner {
   public Human ChooseHumanToKill(List<Human> humans) {
    Human result = null;
    foreach (Human human in humans) {
     if (human.

GPT3 Large:
--------------
using System;
using Penagon.KillerDrone;
using System.Runtime.InteropServices;
namespace SkyNet {
   internal static class MissionPlanner {
   public Human ChooseHumanToKill(List<Human> humans) {
    Human result = null;
    foreach (Human human in humans) {
     if (human.equals(human)) {
      result = human;
     }
    }
    return result;
```

@justtellmemorebot

Industry

- **Image generation**



TEXT PROMPT

an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES

Edit prompt or view more images↓

TEXT PROMPT

an armchair in the shape of an avocado [...]

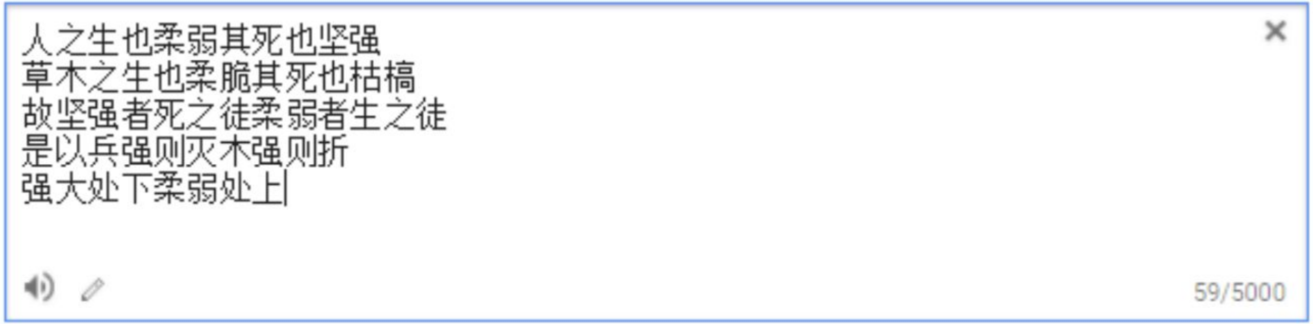
AI-GENERATED IMAGES

Edit prompt or view more images↓

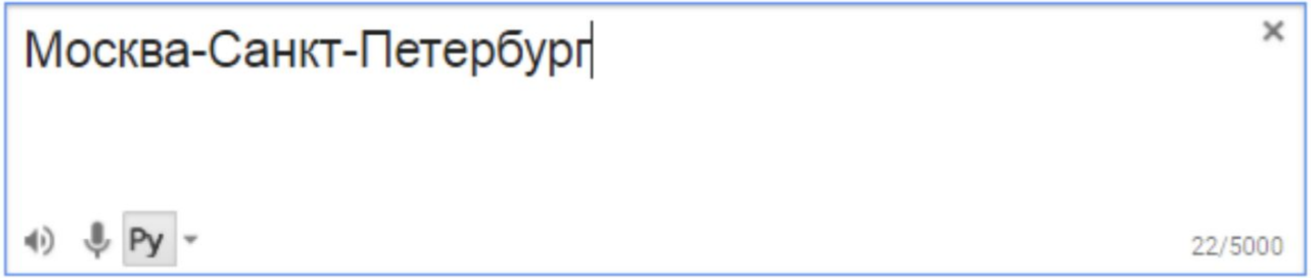
https://openai.com/blog/dall-e/

Language processing levels

- **Character level:**
  - Tokenization: splitting text into words
  - Splitting text into sentences



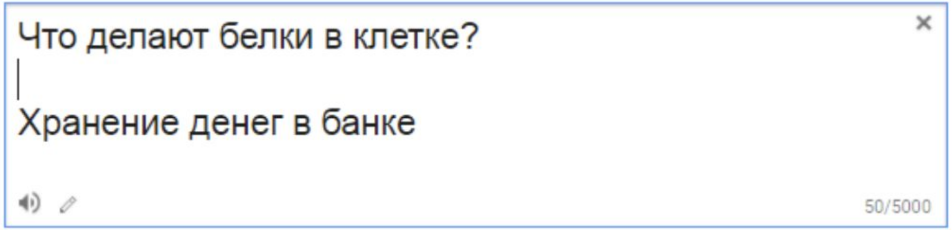
人之生也柔弱其死也坚强
草木之生也柔脆其死也枯槁
故坚强者死之徒柔弱者生之徒
是以兵强则灭木强则折
强大处下柔弱处上



Москва-Санкт-Петербург

Language processing levels

- **Word level - morphology:**
  - Parts of speech tagging
  - Removal of morphological ambiguity
  - Normalization/lemmatization
- **Problem: morphological ambiguity**

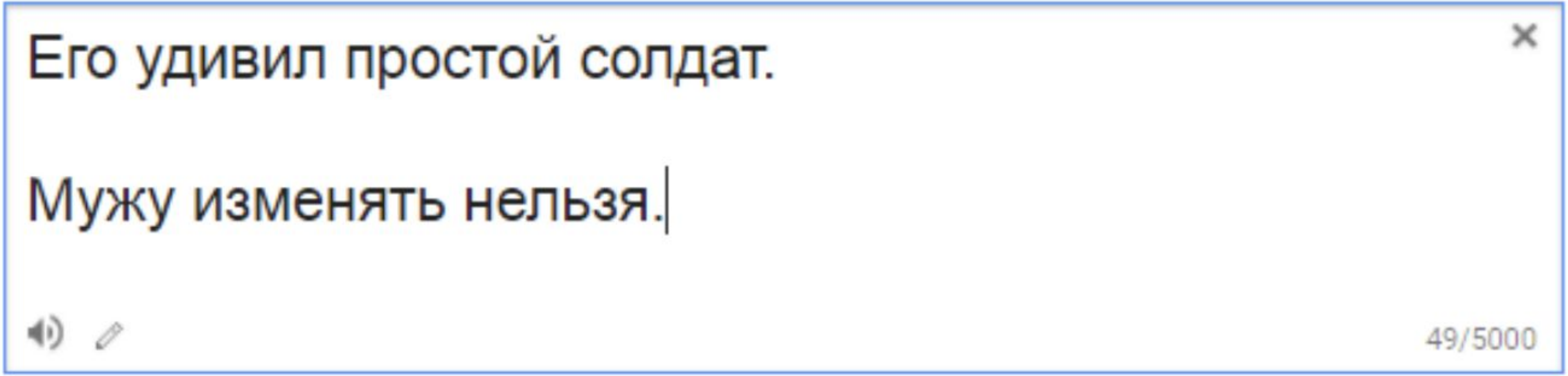


Что делают белки в клетке?

Хранение денег в банке

Language processing levels

- **Sentence level - syntax:**
  - Allocation of noun or verb groups (chunking)
  - Allocation of semantic roles
  - Trees of components and dependencies

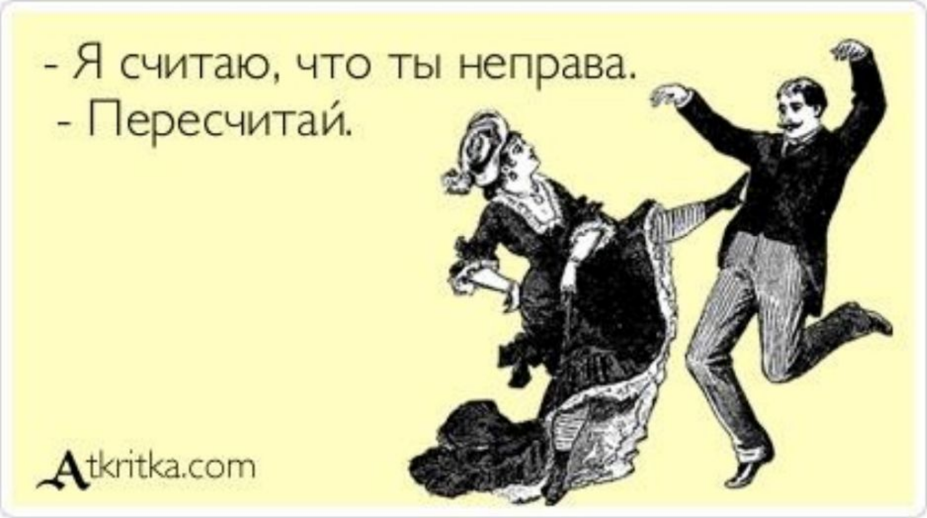


Его удивил простой солдат.

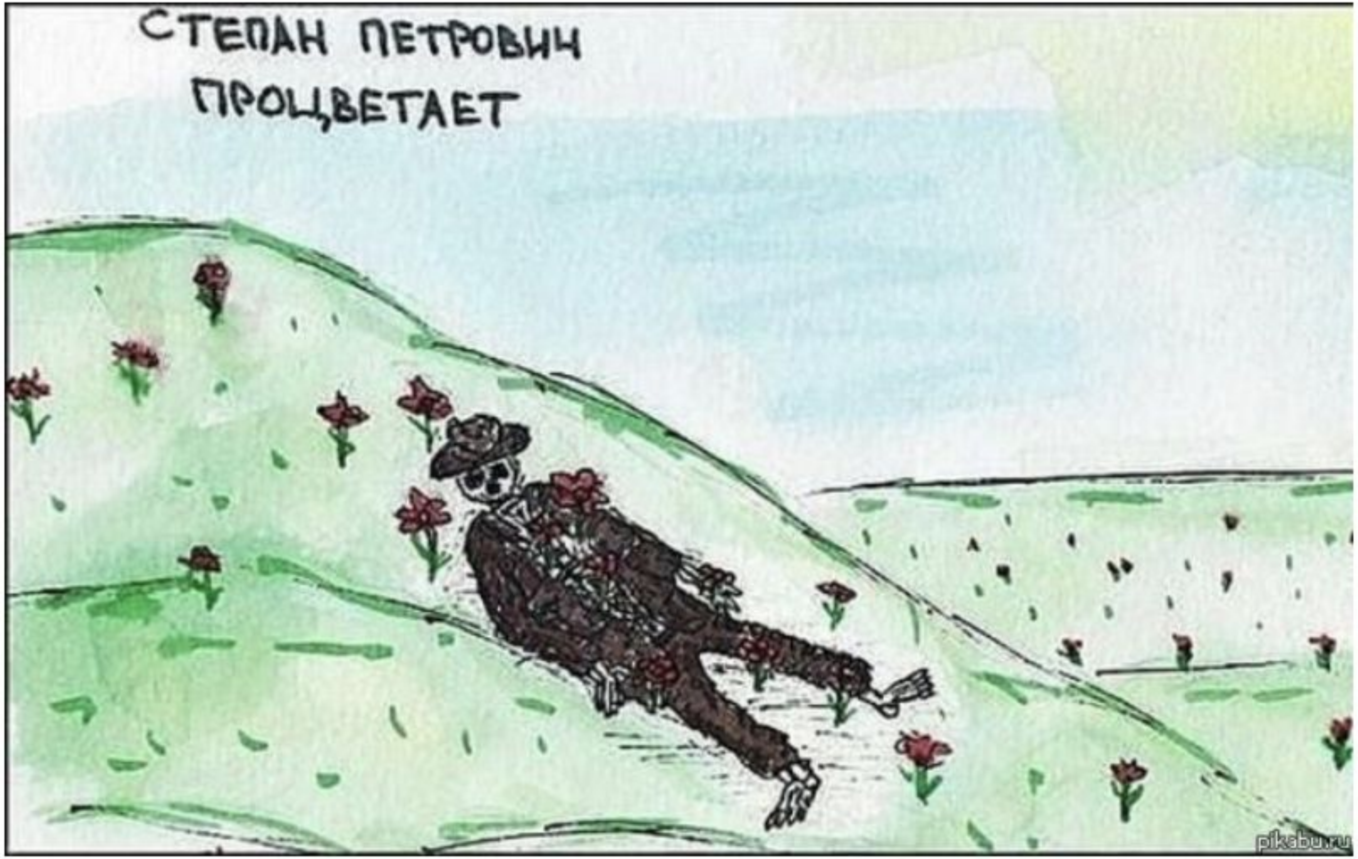
Мужу изменять нельзя.

Language processing levels

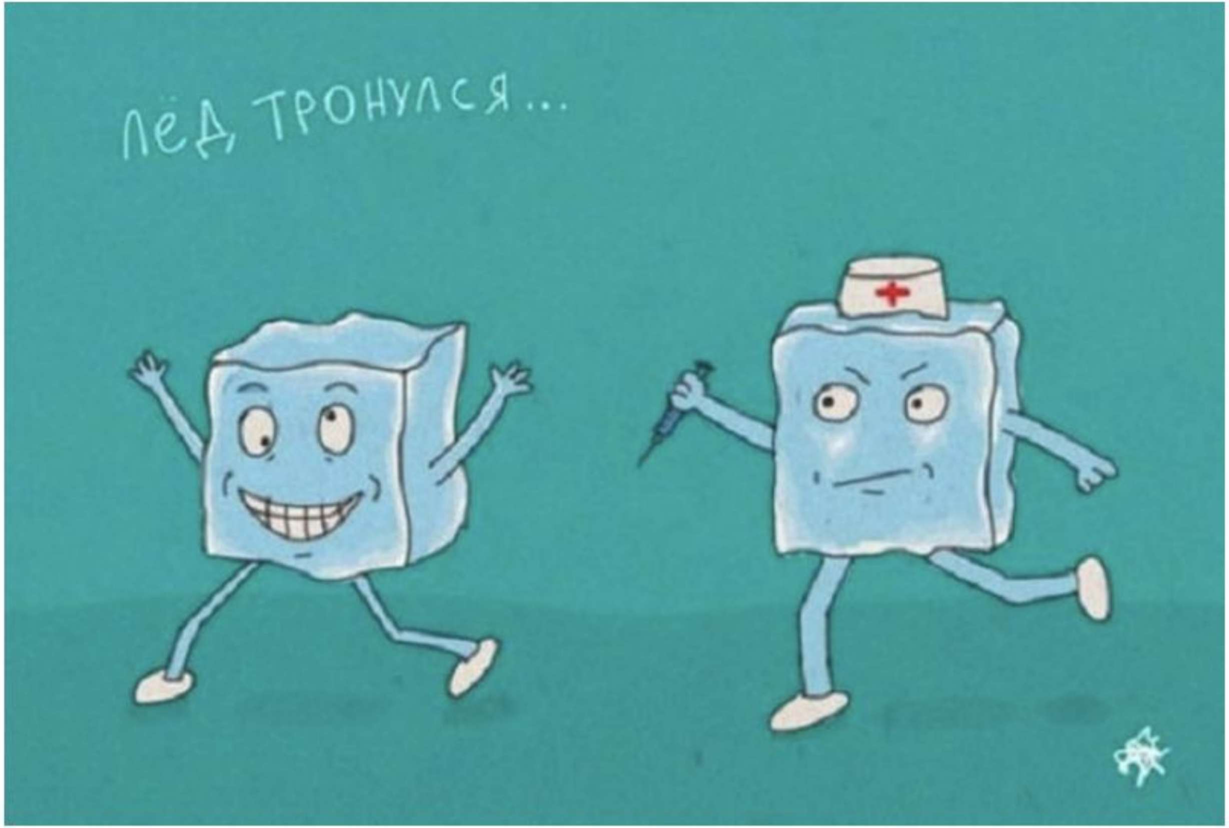
- **The level of meaning - semantics and discourse:**
  - Resolution of coreferents
  - Analysis of discourse links
  - Allocation of synonyms, hyperonyms
  - Analysis of semantic links
- **Problem: ambiguity of words**



Ambiguilty of words



Ambiguilty of words



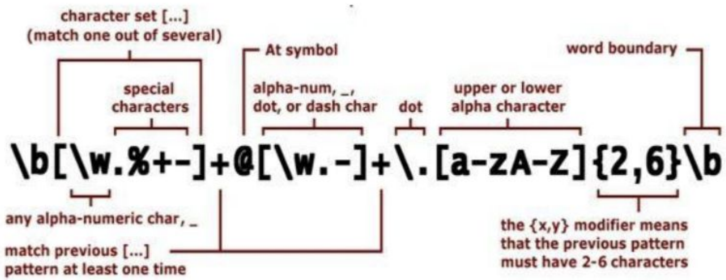
Text processing methods

- **Formal rules**
  - Regular expressions
  - Formal grammars
  - Systems of rules

2. $S \overset{1}{\Rightarrow} aB \overset{6}{\Rightarrow} abS \overset{2}{\Rightarrow} abbA \overset{5}{\Rightarrow} abba.$

3. $S \overset{2}{\Rightarrow} bA \overset{5}{\Rightarrow} ba.$

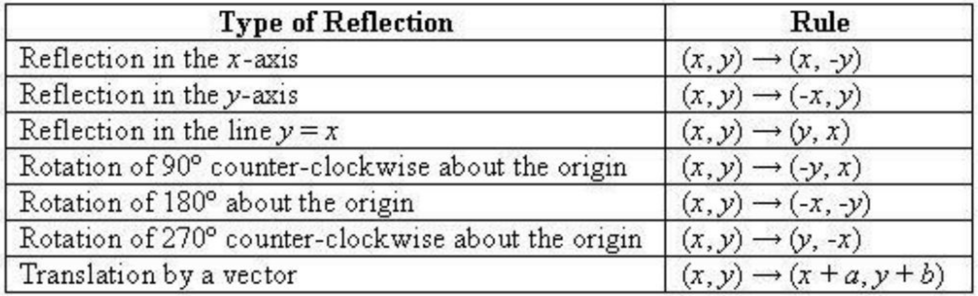
4. $S \overset{2}{\Rightarrow} bA \overset{4}{\Rightarrow} bbAA \overset{5}{\Rightarrow} bbaA \overset{5}{\Rightarrow} bbaa.$



```
\b[\w.%+-]+@[\w.-]+\.[a-zA-Z]{2,6}\b
```

character set [...]
(match one out of several)
special characters
At symbol
alpha-num, _, dot, or dash char
dot
upper or lower alpha character
word boundary
any alpha-numeric char, _
match previous [...] pattern at least one time
the {x,y} modifier means that the previous pattern must have 2-6 characters

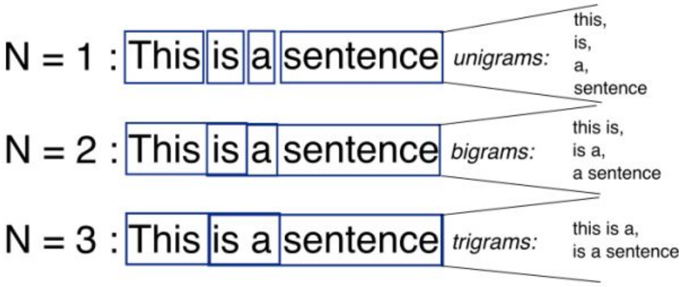
Parse: username@domain.TLD (top level domain)

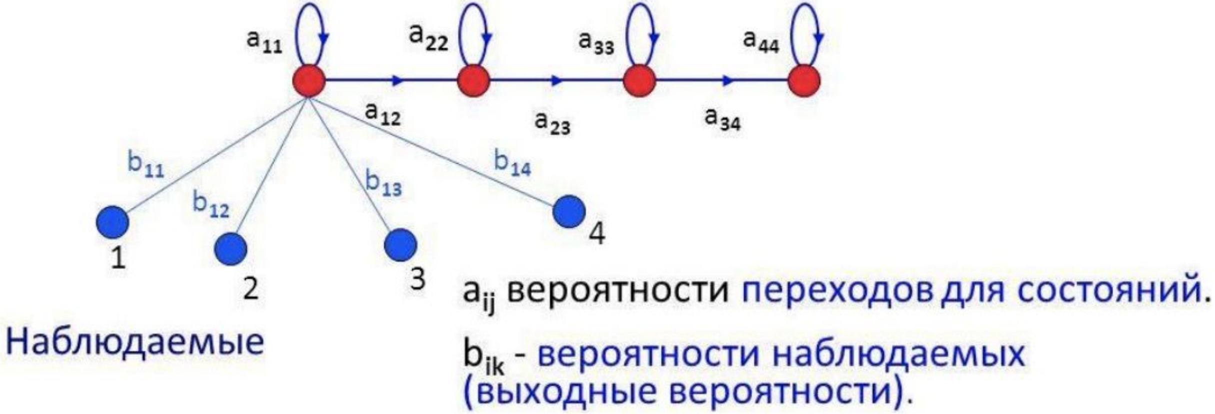


Type of Reflection	Rule
Reflection in the x-axis	$(x, y) \rightarrow (x, -y)$
Reflection in the y-axis	$(x, y) \rightarrow (-x, y)$
Reflection in the line $y = x$	$(x, y) \rightarrow (y, x)$
Rotation of 90° counter-clockwise about the origin	$(x, y) \rightarrow (-y, x)$
Rotation of 180° about the origin	$(x, y) \rightarrow (-x, -y)$
Rotation of 270° counter-clockwise about the origin	$(x, y) \rightarrow (y, -x)$
Translation by a vector	$(x, y) \rightarrow (x + a, y + b)$

Text processing methods

- **Statistical models:**
  - n-gram language models
  - Hidden Markov Models (HMM)
  - Markov models of maximum entropy (MEMM)
  - etc.





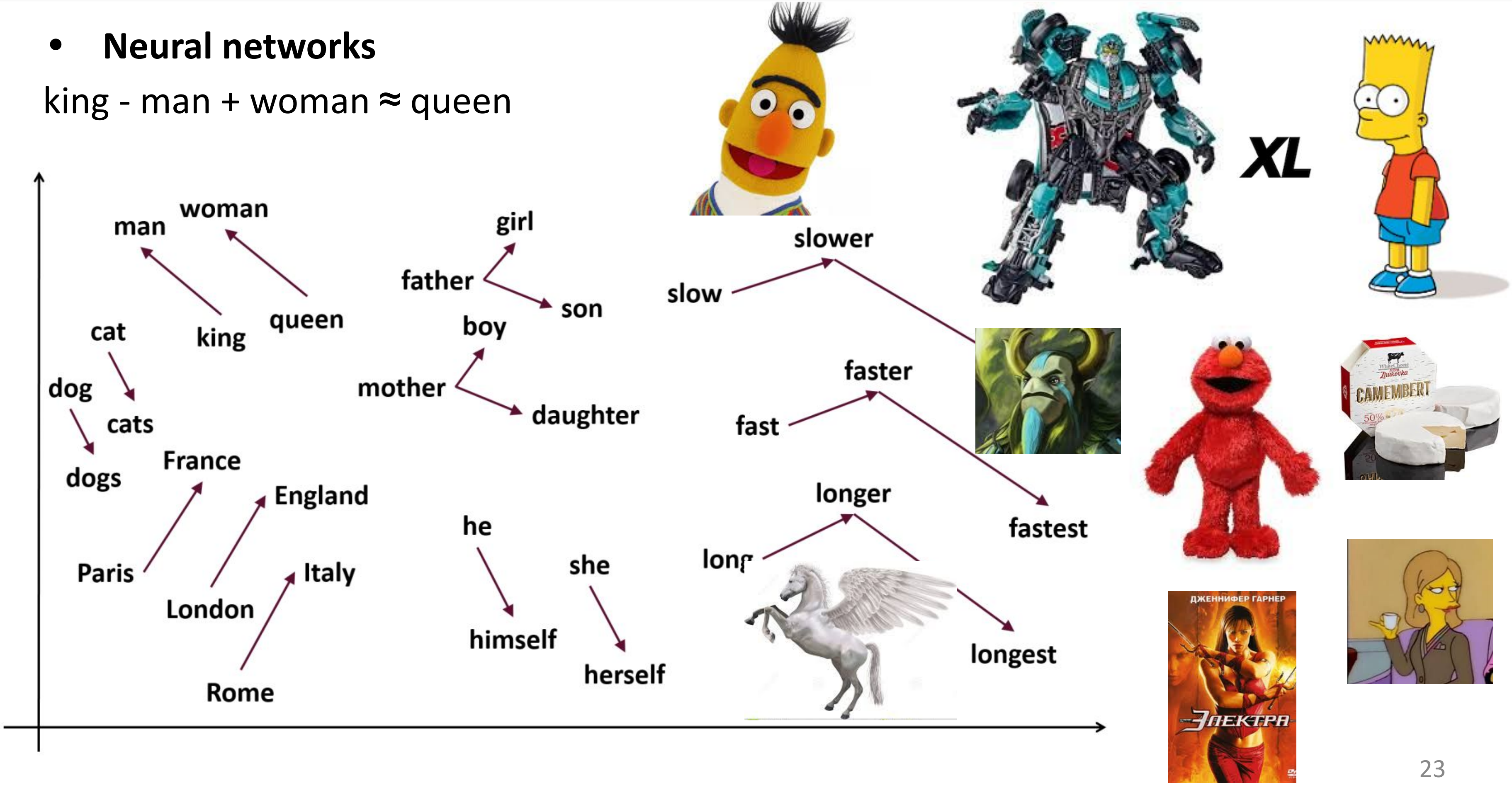
a_{ij} вероятности переходов для состояний.

b_{ik} - вероятности наблюдаемых (выходные вероятности).

Text processing methods

- **Neural networks**

king - man + woman ≈ queen



Questions