

Intro to NLP

MIPT

3.02.2022

Anton Emelianov,
Albina Akhmetgareeva

Today

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

Course instructors



Anton Emelianov, login-const@mail.ru,
[@king_menin](https://www.instagram.com/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>



Albina Akhmetgareeva,
albina.akhmetgareeva@gmail.com,
Tg: [@colindonolwe](https://www.t.me/colindonolwe)

- * Master degree in Computer Science, MIPT
- * R&D NLP, [SberDevices](#), [Sber](#).
- * Linkedin
<https://www.linkedin.com/in/colindonolwe>

Course information

- Github repo: <https://github.com/king-menin/mipt-nlp2022>
- Tlg chat: <https://t.me/+oKrpfmFaGqQ00Dky>
- Final mark: $\text{score}(HWs) / \text{count}(HWs) * 10$ or exam
- Homeworks: ~3, 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	6

- Machine translation

RUSSIAN - DETECTED ENGLISH SPANISH FRENCH ↕ POLISH RUSSIAN ENGLISH

Зачем ты перчишь поросёнка перцем? × Dlaczego pieprzysz prosię? ☆

Zachem ty perchish' porosonka pertsem? 34 / 5000 🔊 🗑️ ✎️ 🔗

- Classification

- Spam filtering
- Sentiment
- Topic or genre

All Mail

Spam (176)

Trash



▼ Categories



👤 **Social (10)**



📧 **Promotions (16)**

📰 **Updates (189)**

💬 **Forums**

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

2  **Темная башня (2017)**
The Dark Tower, WDSFPR
США, реж. Николай Ароэль
(ужасы, фантастика, фантазия...)
Кэтрин Уинник, Мэттью МакКонахи, ...
96% (44 396)  3 августа 2017
Мои фильмы ▼

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

- 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": "000",
    "name": "Комбинат
питания Бутраша"
  },
  "receiving": {
    "type": "ИП",
    "name": {
      "first": "Алёна",
      "middle":
"Александровна",
      "last": "Бугатова"
    }
  },
  "penalties": [
    16085.75,
    597.16,
    2000
  ]
}
```



- Question Answering systems

какая погода в москве

Все Карты Новости Видео Картинки Ещё Настройки Инструменты

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

Температура Вероятность осадков Ветер

 **WolframAlpha** computational intelligence.

what weather in moscow and st.petersburg

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
--	--------	--

Show non-metric More

MTS RUS 54 % 19:00

Ассистент

перевод музыка?

Привет

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

Как дела?


С каждым днем всё лучше. Как дела?

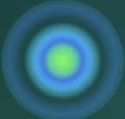
Нормально а жить как


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

Никогда не сдавайся считаешь

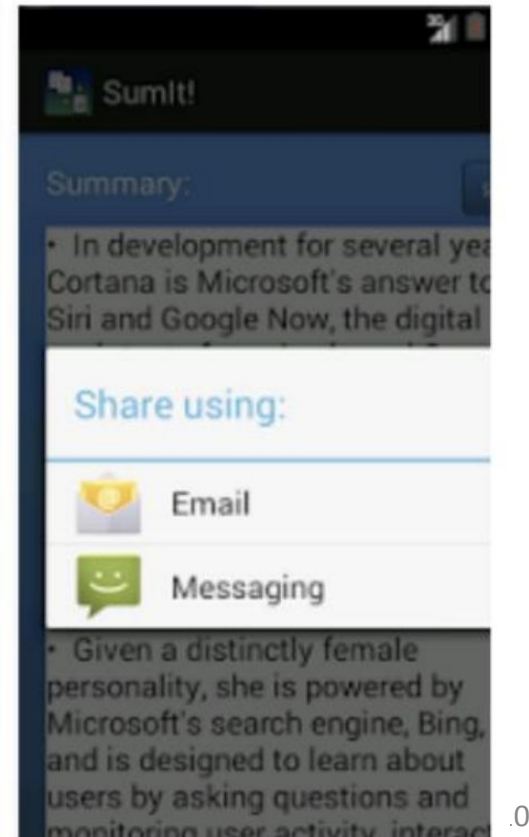
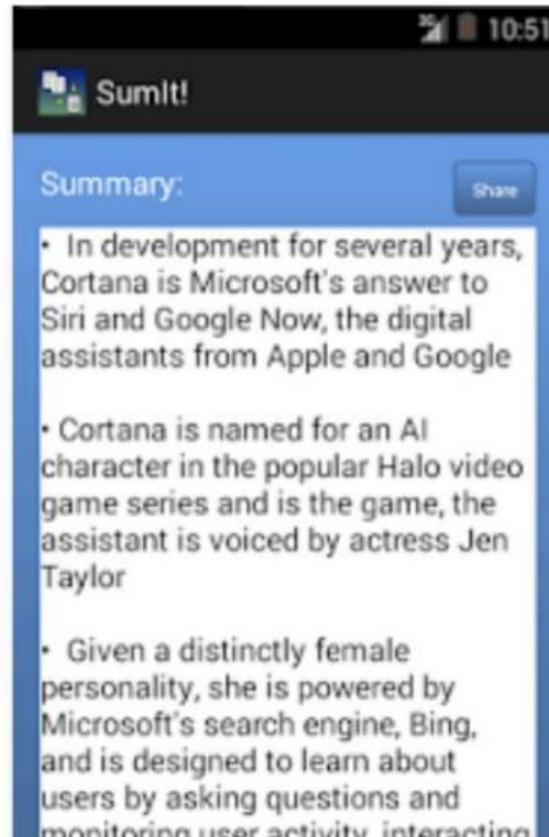
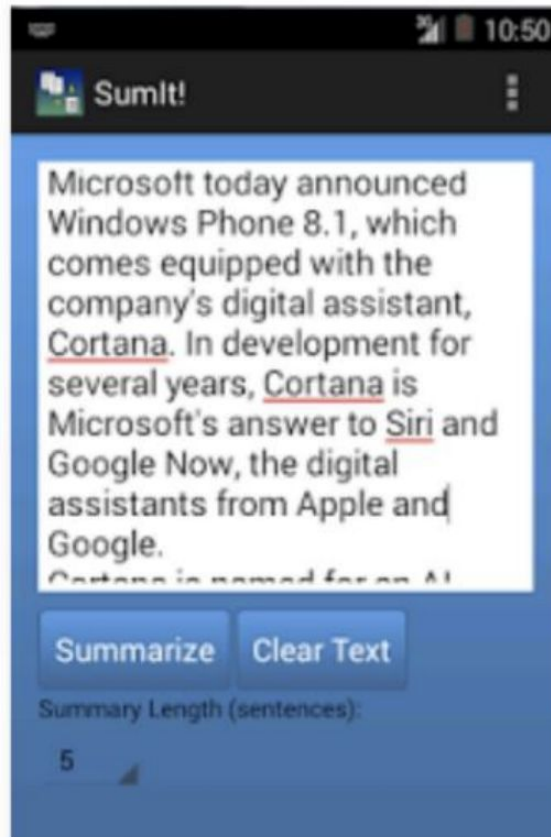
Я всегда стараюсь прикладывать максимальные усилия



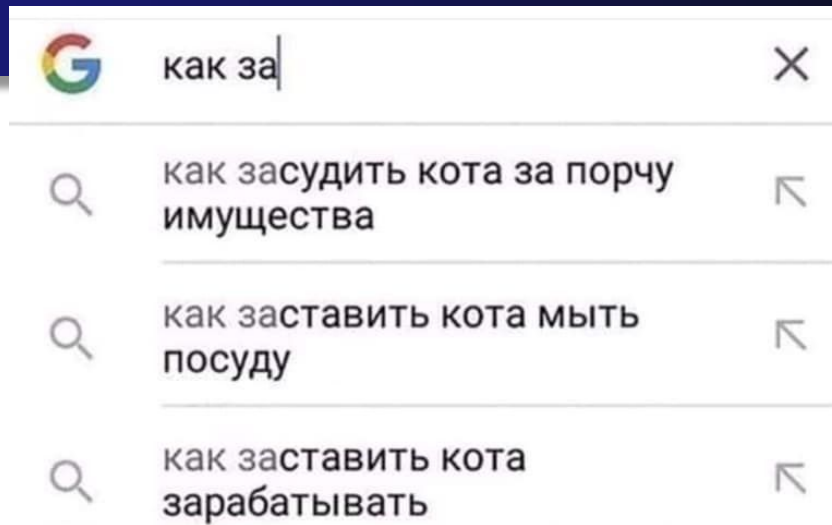




- Text summarization



- Text autocomplete



- Text generation

ruGPT-3 (760M)

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



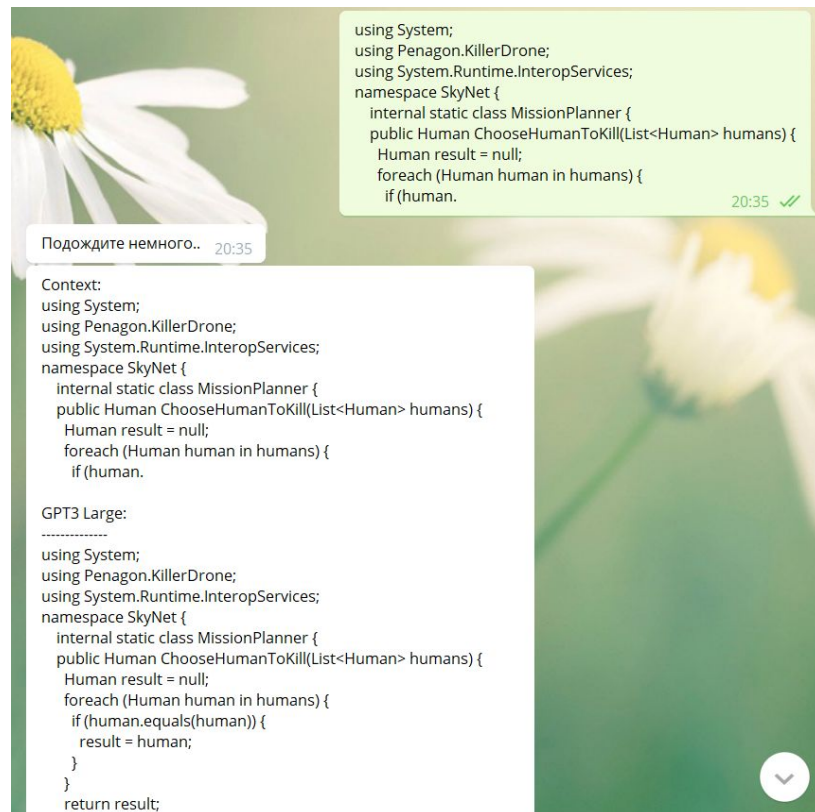
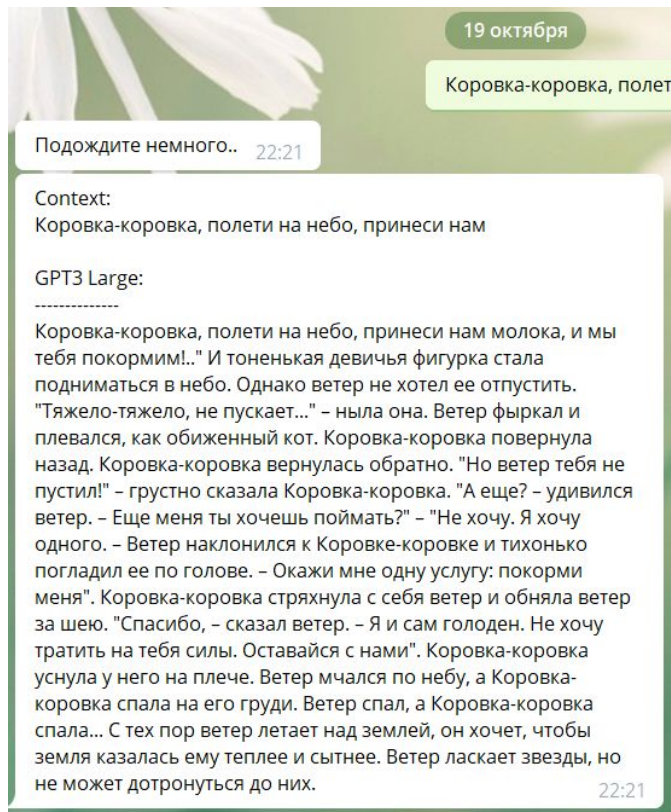
Дополнить

Результат:



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

- Text generation



- 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 [...]

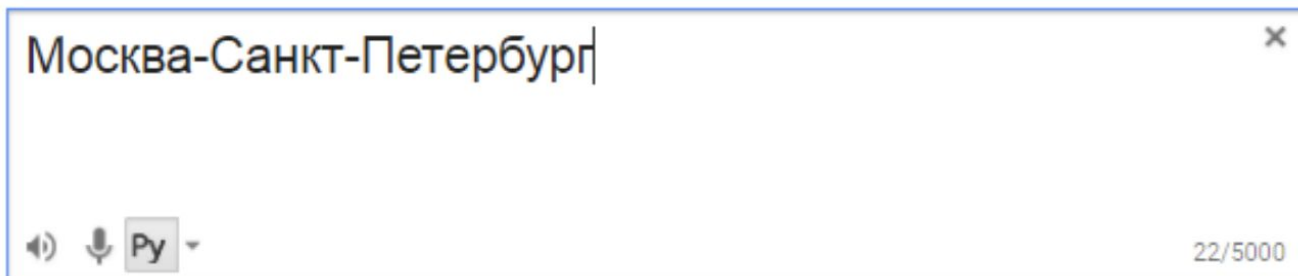
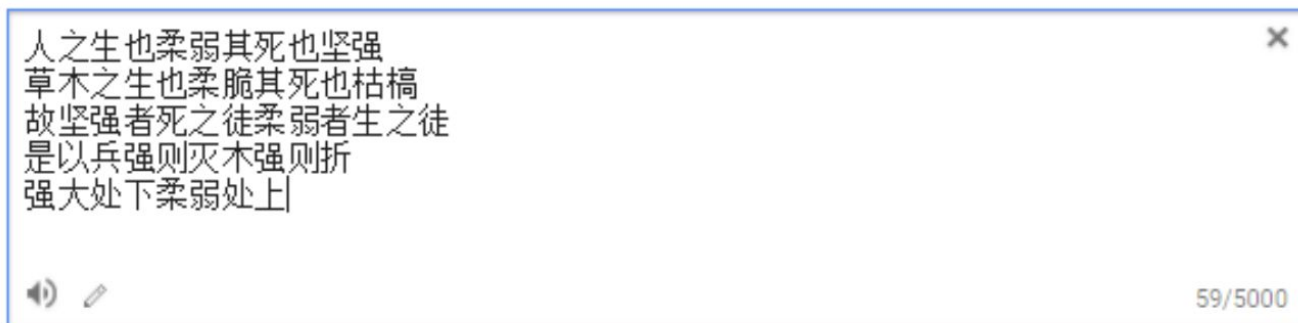
AI-GENERATED IMAGES



Edit prompt or view more images ↓

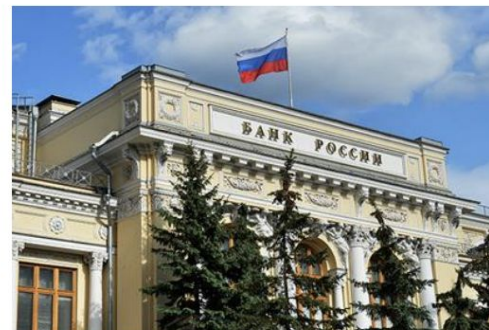
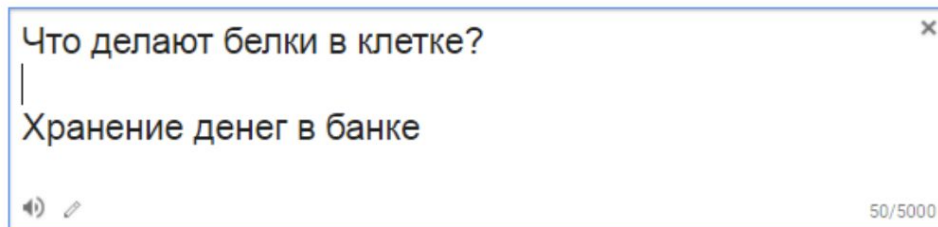
Language processing levels

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



Language processing levels

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



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.



49/5000

Language processing levels

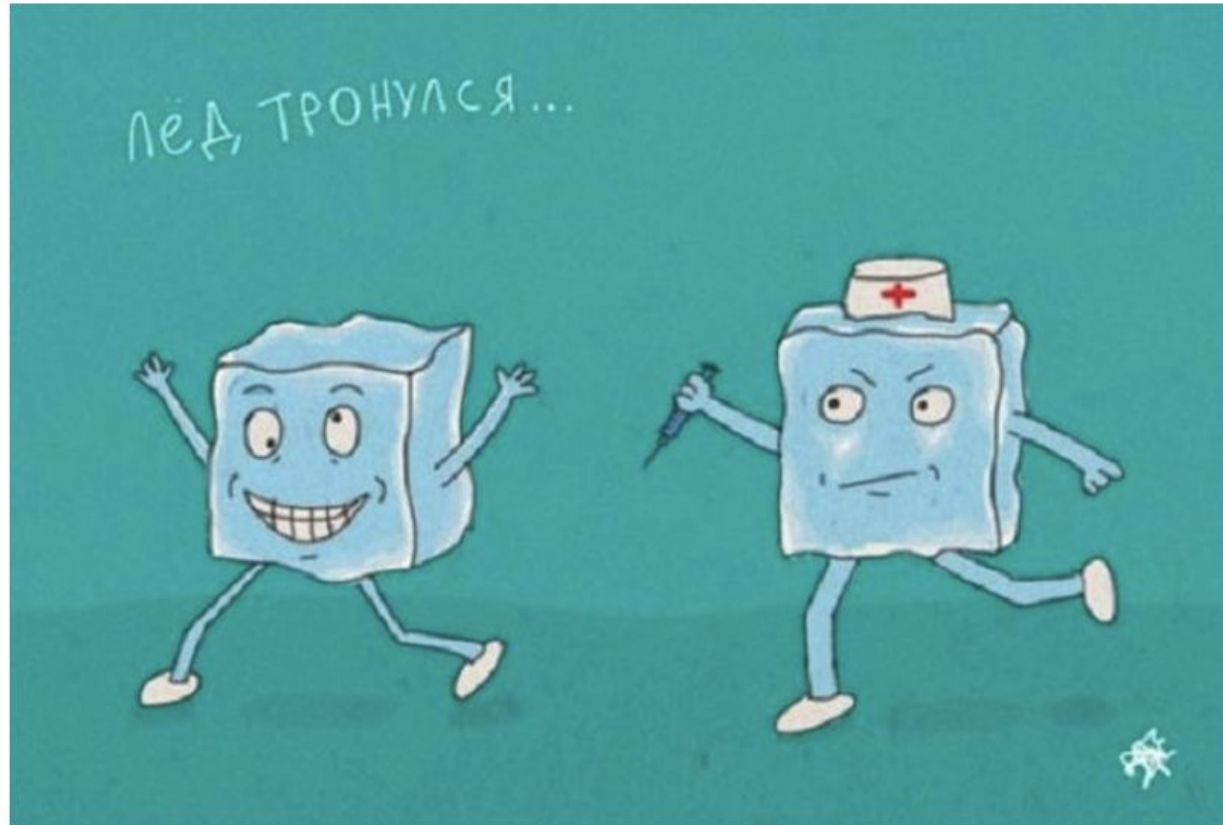
- **The level of meaning - semantics and discourse:**
 - Resolution of coreferents
 - Analysis of discourse links
 - Allocation of synonyms, hyperonyms
 - Analysis of semantic links
- **Problem: ambiguity of words**



Ambiguity of words



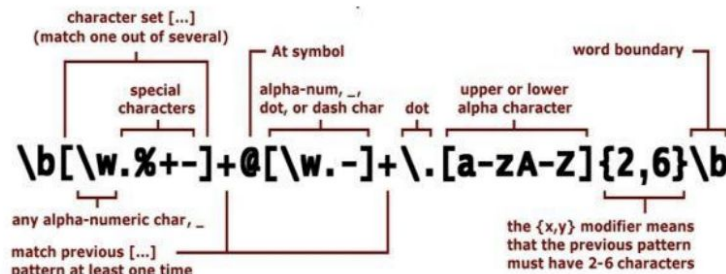
Ambiguity of words



Text processing methods

- Formal rules

- Regular expressions
- Formal grammars
- Systems of rules



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

$$2. \quad S \xRightarrow{1} aB \xRightarrow{6} abS \xRightarrow{2} abbA \xRightarrow{5} abba.$$

$$3. \quad S \xRightarrow{2} bA \xRightarrow{5} ba.$$

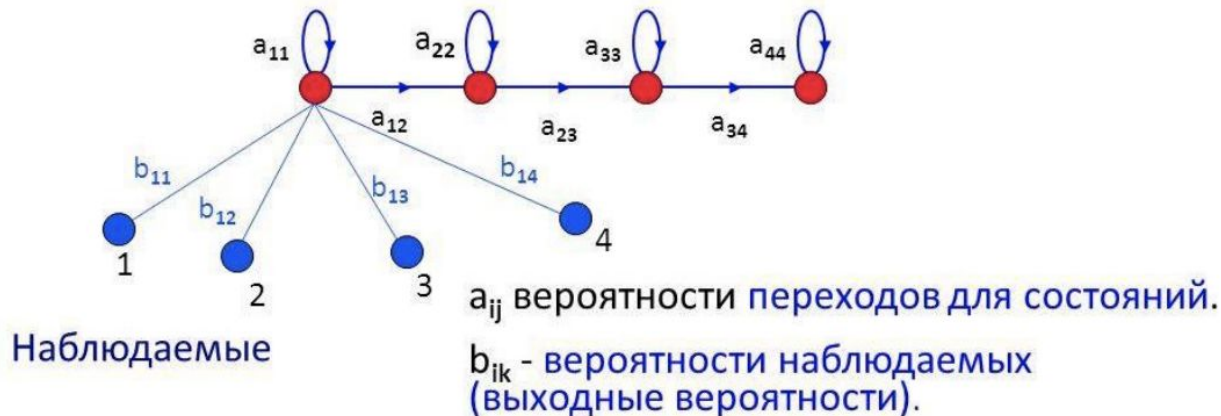
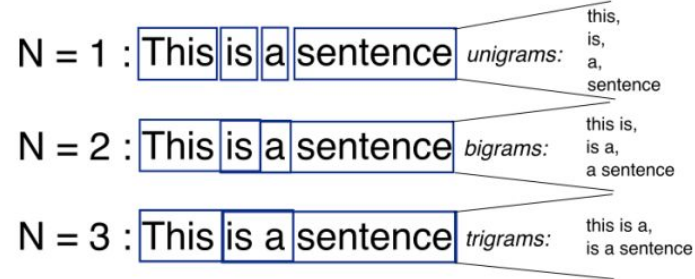
$$4. \quad S \xRightarrow{2} bA \xRightarrow{4} bbAA \xRightarrow{5} bbaA \xRightarrow{5} bbaa.$$

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:**

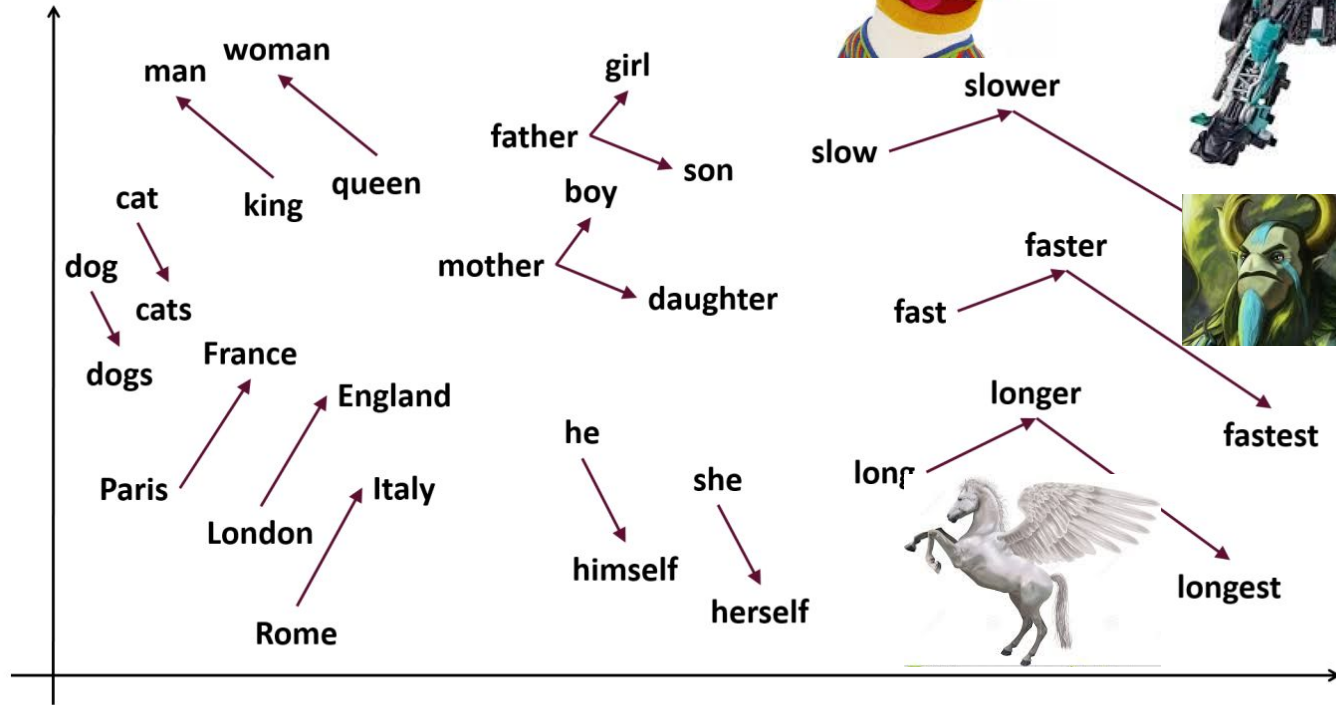
- n-gram language models
- Hidden Markov Models (HMM)
- Markov models of maximum entropy (MEMM)
- etc.



Text processing methods

- Neural networks

king - man + woman \approx queen



XL



Questions