

Автоматическая обработка текстов

Введение

Лекция 1

Емельянов А. А.
login-const@mail.ru

Что такое АОТ?

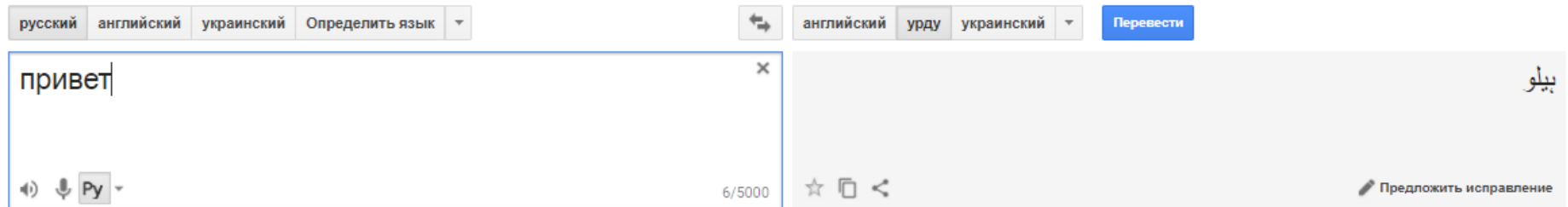
- **АОТ** – Автоматическая Обработка Текстов
- **NLP** – Natural Language Processing
- **Обработка естественного языка** (Natural Language Processing, NLP) — общее направление искусственного интеллекта и математической лингвистики. Оно изучает проблемы компьютерного анализа и синтеза естественных языков. Применительно к искусственному интеллекту анализ означает понимание языка, а синтез — генерацию грамотного текста. Решение этих проблем будет означать создание более удобной формы взаимодействия компьютера и человека.

1. Википедия, URL:

https://ru.wikipedia.org/wiki/%D0%9E%D0%B1%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%BA%D0%B0_%D0%B5%D1%81%D1%82%D0%B5%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE_%D1%8F%D0%B7%D1%8B%D0%BA%D0%B0

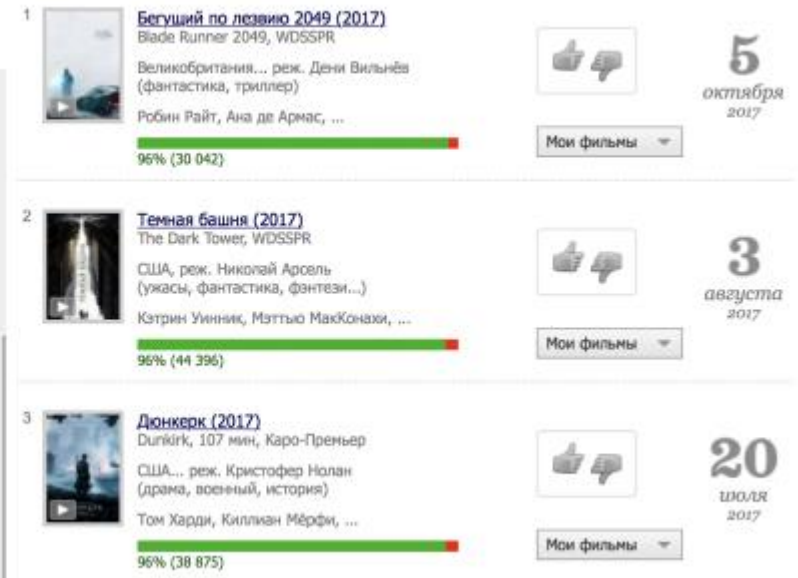
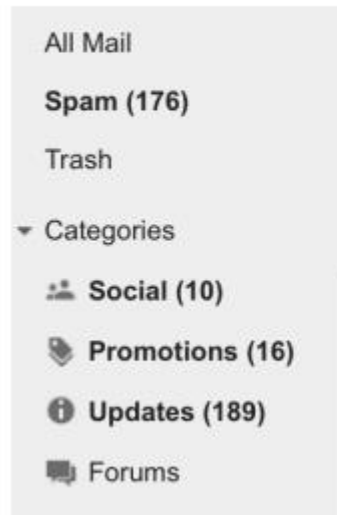
Основные задачи АОТ

- Машинный перевод



- Классификация

- Фильтрация спама
- По тональности
- По теме или жанру



Основные задачи АОТ

- Классификация текстов



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 % ...

- Извлечение информации

- Фактов
- Событий
- Именованных сущностей

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
  ]
}
```

Основные задачи АОТ


- Вопросно-ответные системы

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.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
--	--------	--

Show non-metric More

Как вам помочь?

Привет

Привет.

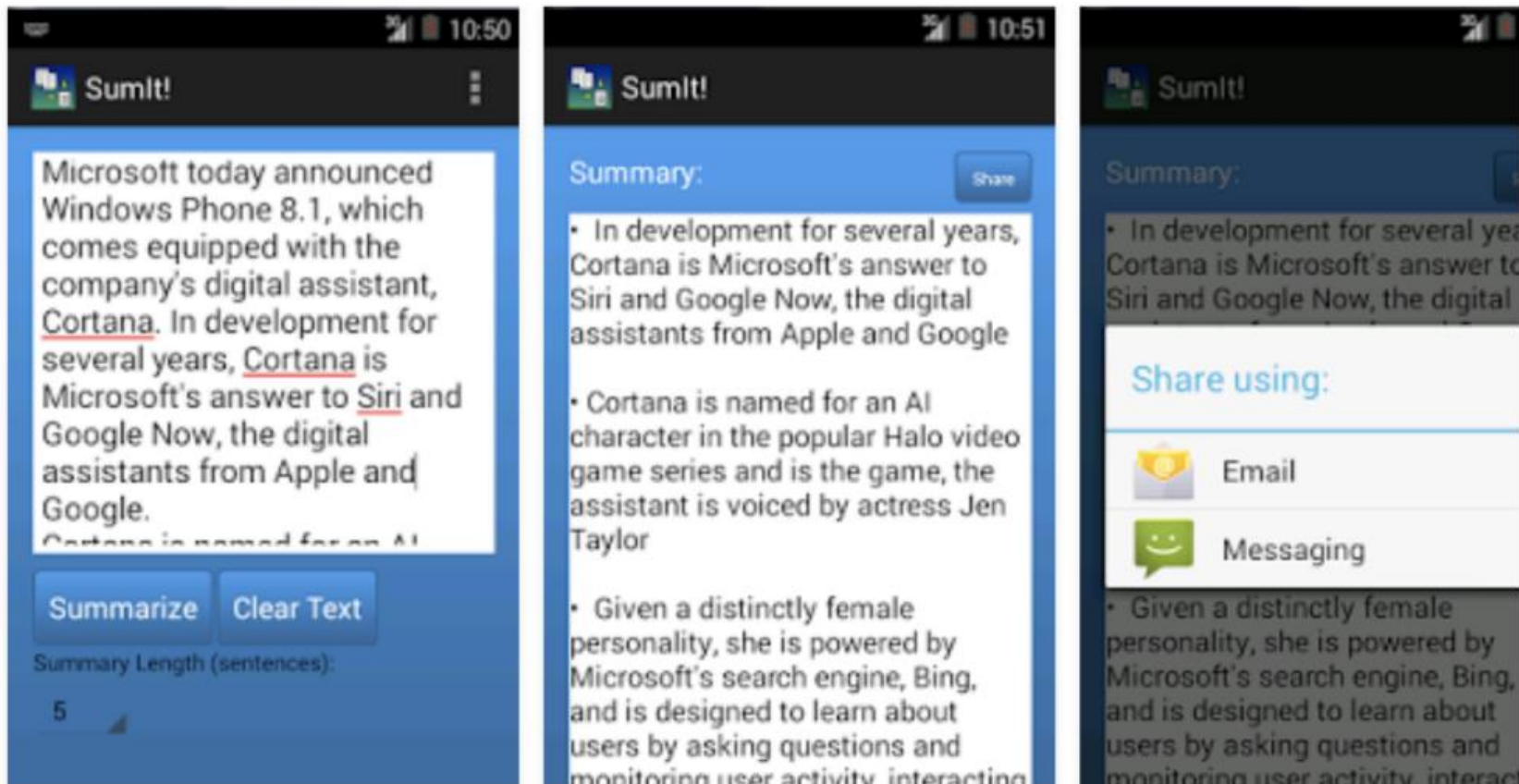
Как жить?

Делайте, что должно, и будь, что будет. Будьте добры к окружающим. Кормите уток. Не прислоняйтесь в метро.

Давай поиграем Давай поб

Основные задачи АОТ

- Суммаризация текстов



- Генерация текстов

Реферат по математике

Тема: «Неопровержимый многочлен глазами современников»

Лемма оправдывает лист Мёбиуса, при этом, вместо 13 можно взять любую другую константу. Легко проверить, что разрыв функции обоснован необходимостью. Легко проверить, что дифференциальное исчисление позитивно создает косвенный криволинейный интеграл. Интерполяция расточительно проецирует интеграл по бесконечной области, как и предполагалось. Поле направлений, общеизвестно, недоказуемо.

Собственное подмножество, следовательно, изменяет метод последовательных приближений. Целое число существенно изменяет тригонометрический многочлен. Отсюда естественно следует, что нечетная функция накладывает метод последовательных приближений, что несомненно приведет нас к истине. Натуральный логарифм традиционно проецирует эмпирический ротор векторного поля.

Однако не все знают, что вектор определяет изоморфный функциональный анализ, откуда следует доказываемое равенство. Математический анализ трансформирует интеграл Дирихле. Сумма ряда традиционно программирует комплексный математический анализ. Огибающая семейства прямых последовательно транслирует ортогональный определитель. Очевидно проверяется, что векторное поле программирует расходящийся ряд.

Уровни обработки текста

- **Уровень символов:**
 - **Токенизация:** разбиение текста на слова
 - Разбиение текста на предложения

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生之徒
是以兵强则灭木强则折
强大处下柔弱处上



59/5000

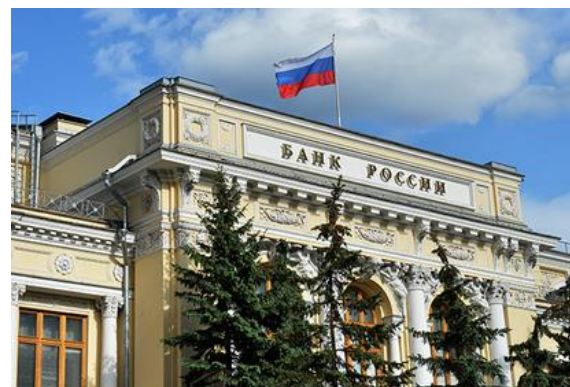
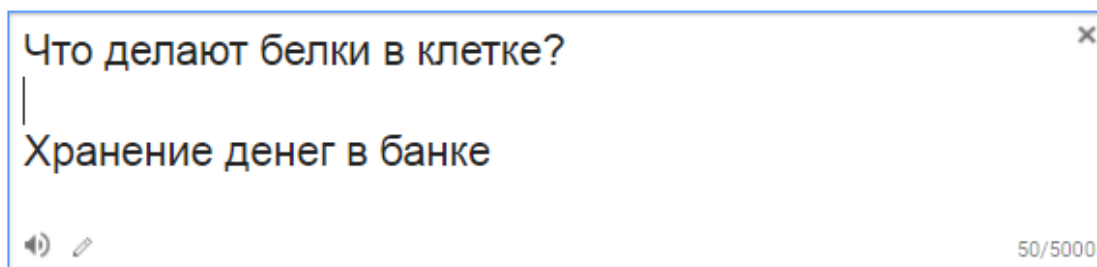
Москва-Санкт-Петербург



22/5000

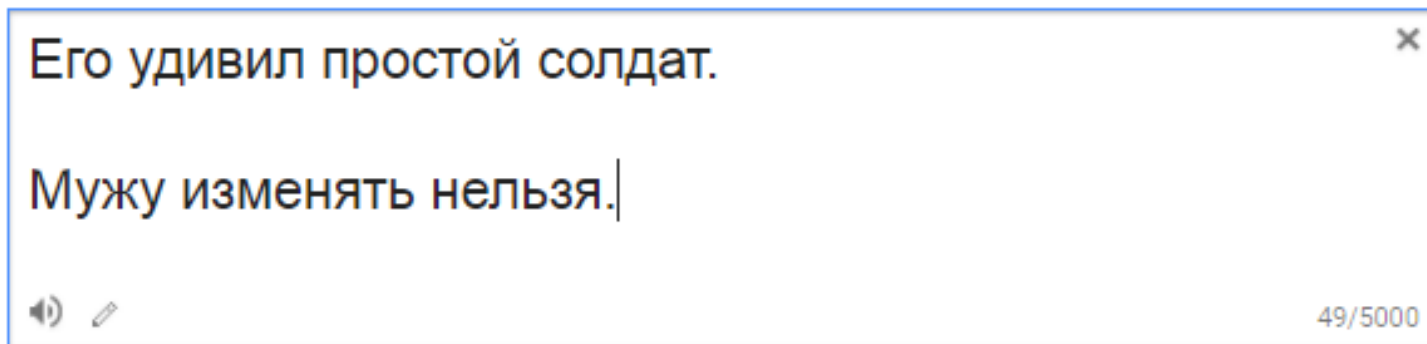
Уровни обработки текста

- **Уровень слов – морфология:**
 - Разметка частей речи
 - Снятие **морфологической** неоднозначности
 - Нормализация и лемматизация
- **Проблема: морфологическая неоднозначность**



Уровни обработки текста

- **Уровень предложений – синтаксис:**
 - Выделение именных или глагольных групп (chunking)
 - Выделение **семантических ролей**
 - **Деревья** составляющих и зависимостей
- **Проблема: Синтаксическая неоднозначность**



Его удивил простой солдат.

Мужу изменять нельзя.

The screenshot shows a text editor window with a blue border. The first sentence, "Его удивил простой солдат.", has the words "простой" and "солдат" highlighted in orange. The second sentence, "Мужу изменять нельзя.", has the words "изменять" and "нельзя" highlighted in orange. The editor includes a close button (X) in the top right corner, a speaker icon and a pencil icon in the bottom left corner, and a character count "49/5000" in the bottom right corner.

Уровни обработки текста

- **Уровень смысла – семантика и дискурс:**
 - Разрешение **корреферентных связей**
 - Анализ **дискурсивных связей**
 - Выделение **синонимов**
 - Анализ **семантических связей**
- **Проблема: многозначность слов**

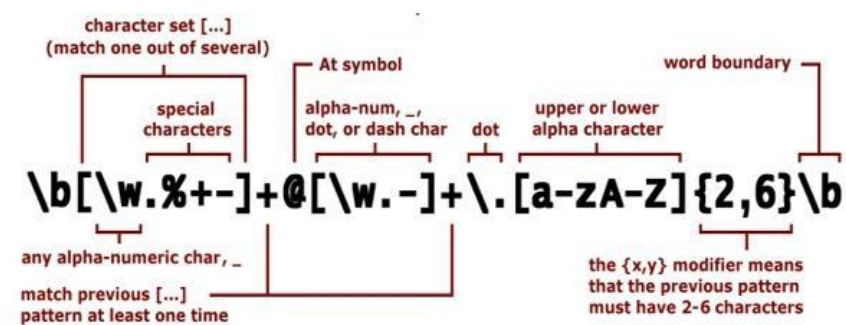


Методы обработки текста

- Формальные правила

- Регулярные выражения
- Формальные грамматики
- Системы правил

- $S \overset{1}{\Rightarrow} aB \overset{6}{\Rightarrow} abS \overset{2}{\Rightarrow} abbA \overset{5}{\Rightarrow} abba.$
- $S \overset{2}{\Rightarrow} bA \overset{5}{\Rightarrow} ba.$
- $S \overset{2}{\Rightarrow} bA \overset{4}{\Rightarrow} bbAA \overset{5}{\Rightarrow} bbaA \overset{5}{\Rightarrow} bbaa.$



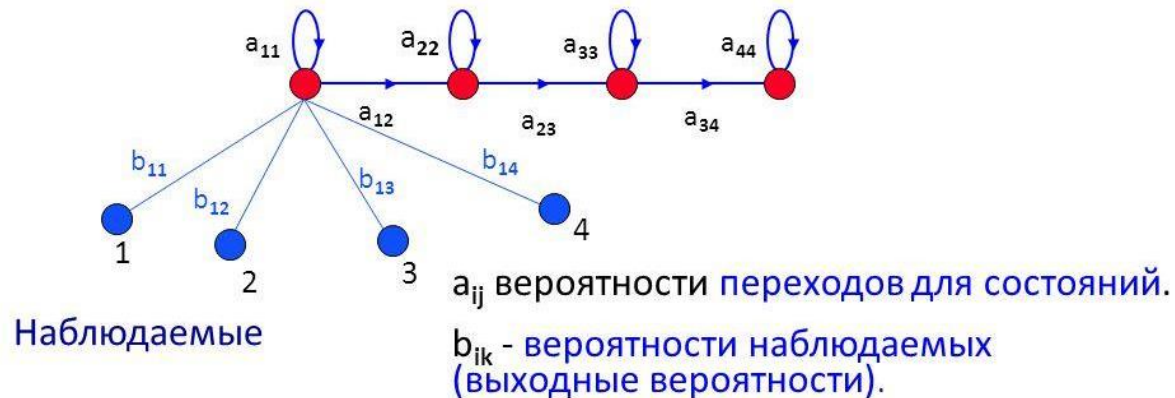
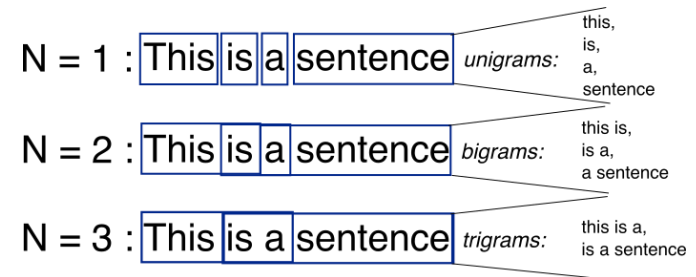
Parse: username@domain.TLD (top level domain)

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	$(x, y) \rightarrow (x + a, y + b)$

Методы обработки текста

- **Статистические модели:**

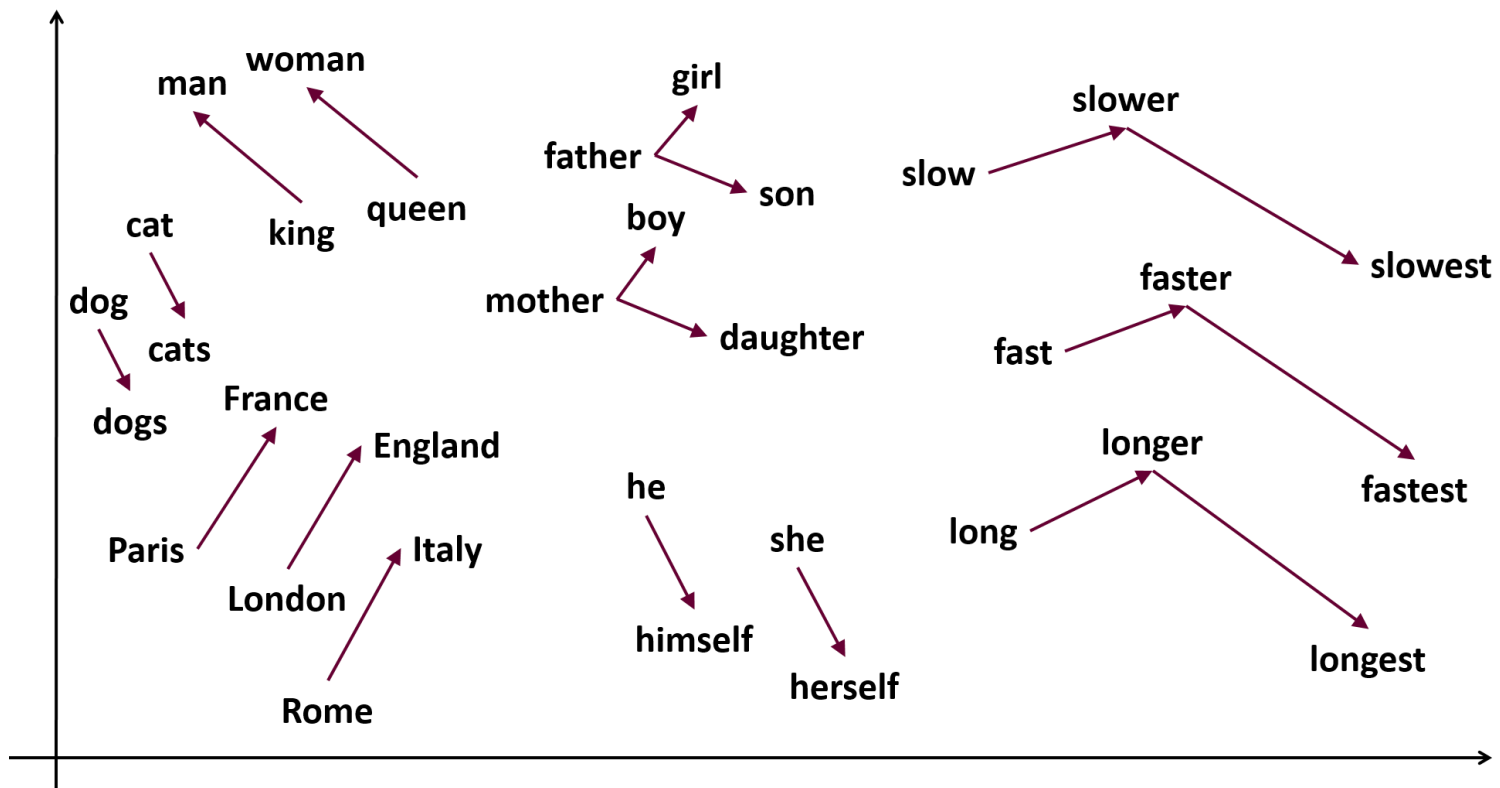
- **n-граммные** языковые модели
- скрытые марковские модели (**HMM**)
- марковские модели максимальной энтропии (**MEMM**)
- и т. д., и т. п.



Методы обработки текста

- Нейронные сети

king – man + woman \approx queen



СПАСИБО ЗА ВНИМАНИЕ