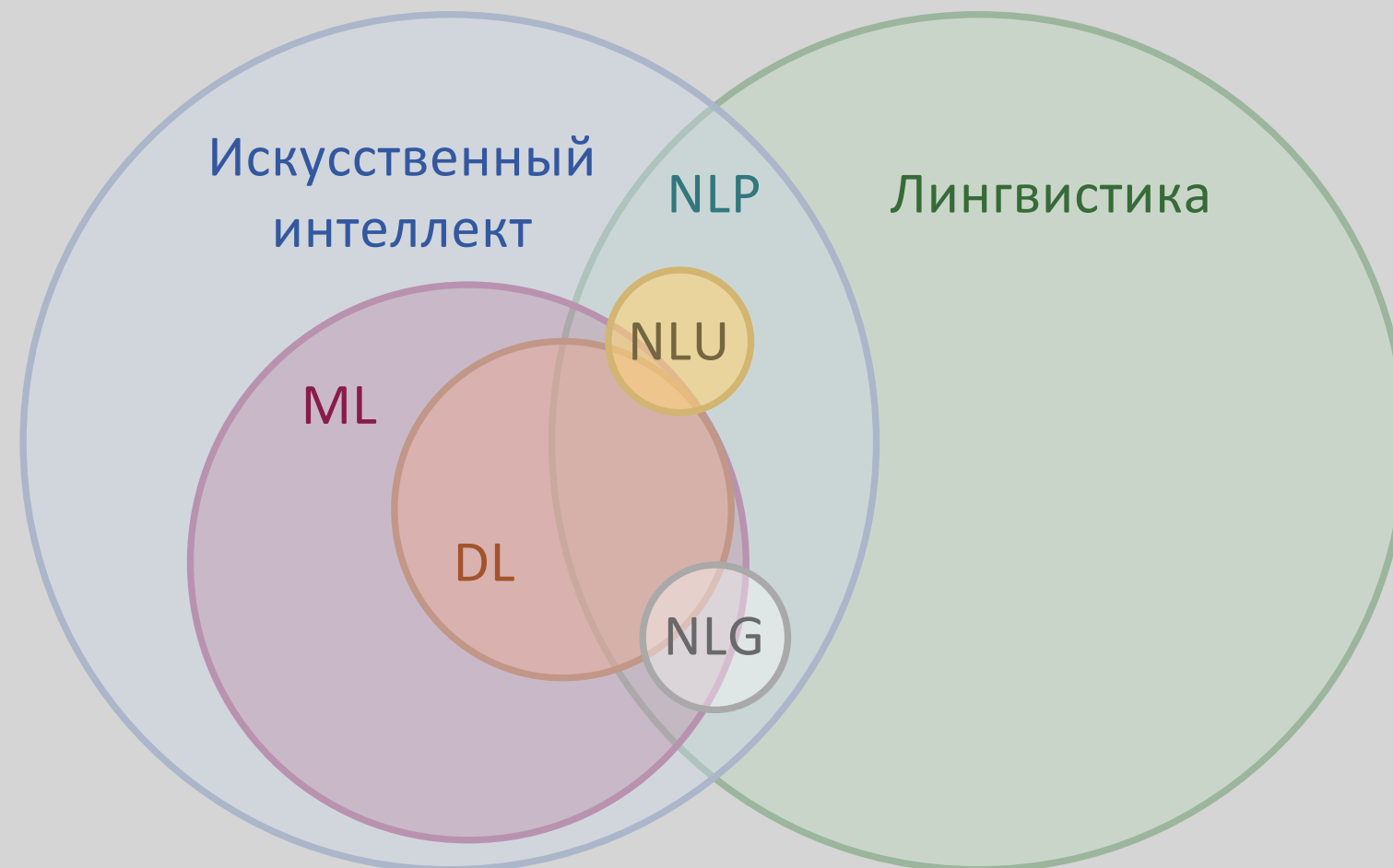


Глубинное обучение для текстовых данных

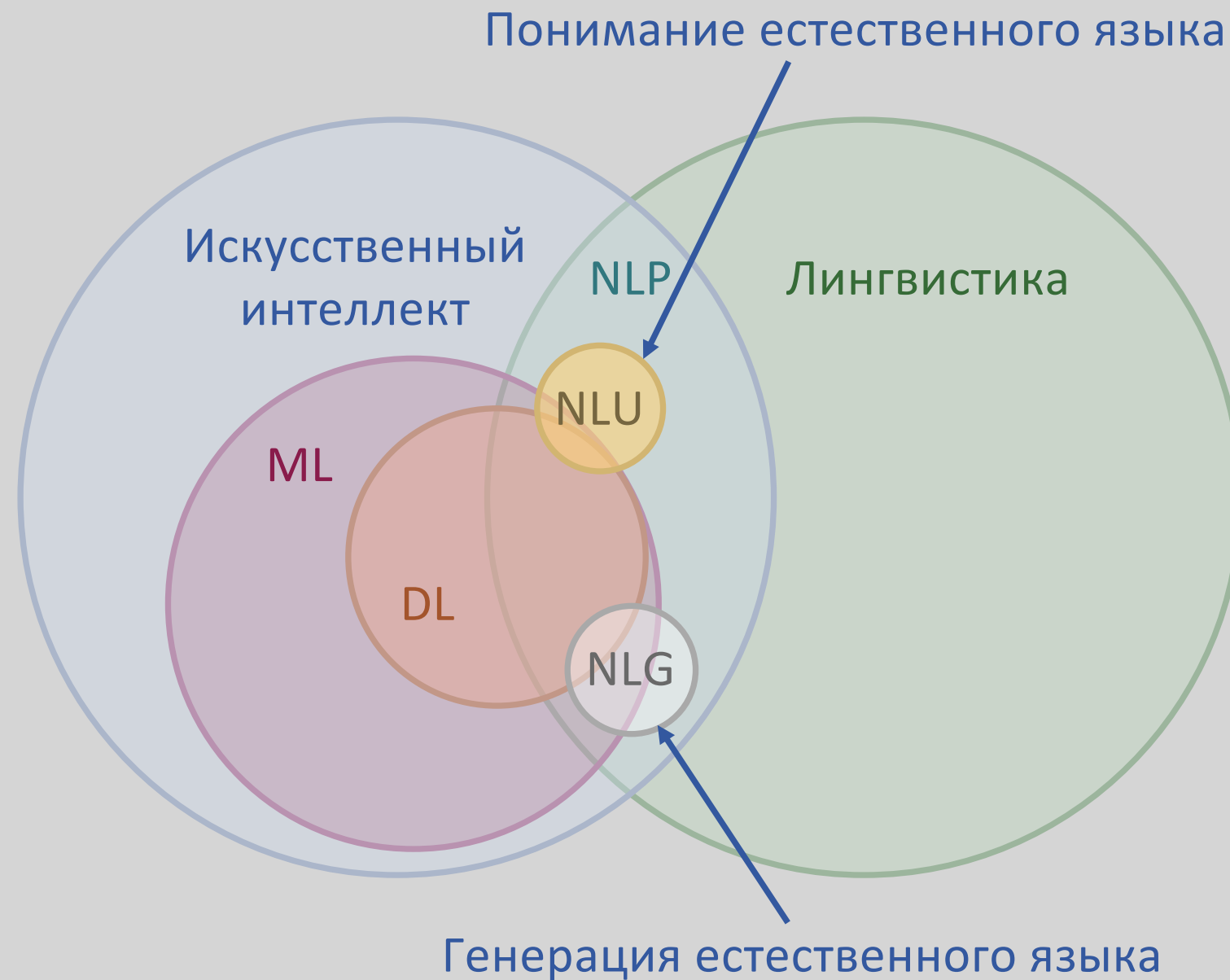
I. Введение

Екатерина Артемова (на основе лекций М. Апишева, Д. Кирьянова)

Автоматическая обработка текстов



Автоматическая обработка текстов



Языковые технологии



Почему работать с текстами сложно?

Многозначность

*Он был в отличной
форме, но на
животе она уже не
застегивалась.*

Сленг

*Жиза
Зашквар
Кекнуться
Каеф*

Идиомы

*Мир тесен
Вышла из себя
Играть с огнем
Не месяц май*

Омонимия

*Стали
Стекло
Косой*

Неологизмы

*Постить
Зафрендить
Каршеринг*

Нестандартный язык

*ем можед личку
прочетаеш*

Почему работать с текстами сложно?

Знание об окружающем мире

Даша [обругала / обняла] Машу, потому что она была расстроена.

Кто расставивался? [Даша / Маша]

Сложные именованные сущности

«Анна Каренина» стала первым российским мюзиклом

Данные

Разметка данных – трудоемкий и дорогой процесс.

Почему работать с текстами сложно?

- Слово – базовая структурная единица языка
- Даже без контекста слово само по себе несет много полезной информации
- Слов очень много, поэтому возникают **большие и разреженные признаковые пространства**

Почему работать с текстами сложно?

- Язык определяет внутреннюю структуру:
 - Слова и словосочетания (морфология)
 - Предложения (синтаксис)
 - Текст (дискурс, порядок изложения)

Почему работать с текстами сложно?

Большое число сырых текстов

Наличие в них языковой структуры



Обучение **больших общезыковых моделей** на сырых данных

Содержание курса

Часть 1. Базовый уровень

- Модели представления текста: векторная модель, языковая модель
- Уровни анализа текста: морфологолический анализ, синтаксический анализ
- Основные постановки задач: классификация текстов, извлечение именованных сущностей, машинный перевод

Часть 2. Продвинутый уровень

- Модели представления текста: контекстуализированные модели, тематическое моделирование, графы знаний,
- Постановки задач: суммаризация текстов, вопросно-ответные системы, поисковые системы

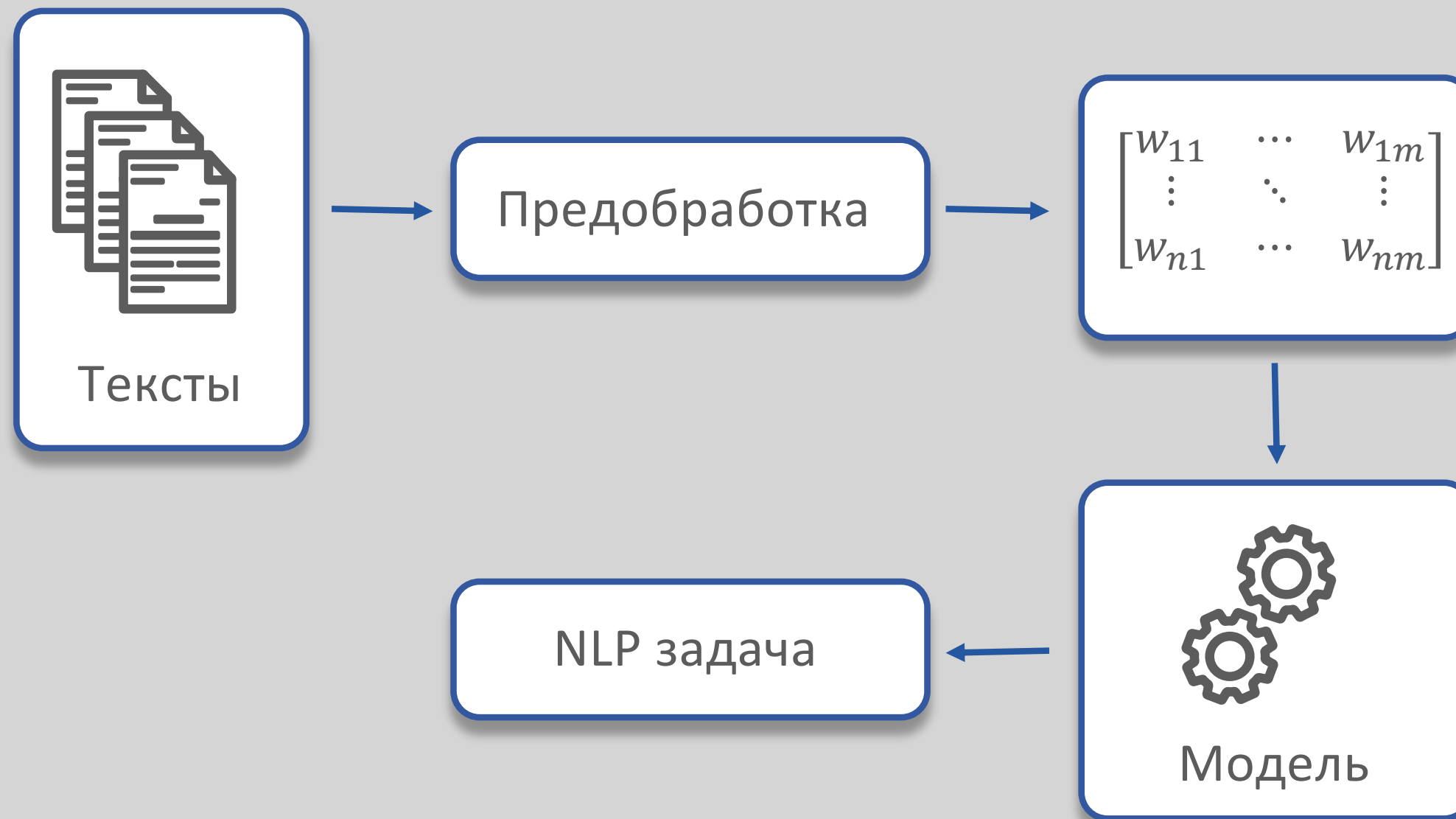
Структура курса

ФОРМУЛА ОЦЕНКИ:

0.3 домашка + 0.2 квизы + 0.5 экз

Роль машинного обучения в обработке текстов

Знание о языке + знание об окружающем мире + данные + машинное обучение = (почти всегда) успех!



Что нужно знать?

- Основные понятия теории вероятностей и математической статистики и линейной алгебры
- Базовые алгоритмы машинного обучения
- Простые нейросетевые модели
- Программировать на Python
- Как использовать iPython-блокноты

Что вы узнаете?

Вы научитесь решать основные задачи обработки текстов.

- Основные понятия автоматической обработки текстов
- Практические постановки задач, возникающие при работе с текстовыми данными
- Архитектуры нейронных сетей, используемые при работе с текстовыми данными
- Инструменты для работы с русским и английским

Индустриальные задачи обработки текстов

Классификация текстов

Большой спектр промышленных задач

- Фильтрация спама
- Анализ тональности
- Категоризация новостей и сообщений
- Выявление недоброкачественных отзывов

Никогда больше не стану
покупать у вас пончики. В
пончике должна быть ОДНА
дырка посередине, а в вашем
пончике я нашла 15! Ну это
никуда не годится!

Обожаю пончики на завтрак!
Это лучшее начало трудового
дня, они заряжают энергией и
хорошим настроением меня и
всю мою семью. Гомер С.

Ранжирование

Сортировка текстов в соответствии с некоторыми критериями

- Информационный поиск – релевантность текста пользовательскому запросу
- Рекомендательные системы – близость интересам пользователя

Машинный перевод

Перевести текст с одного языка на другой

The shop owner caught the boy red-handed
when he was stealing cigarettes.



Хозяин магазина поймал парня
красноруким, когда он воровал
сигареты.



Хозяин магазина поймал парня с
поличным, когда он воровал
сигареты.

Анализ и коррекция текста

Исправление опечаток, ошибок согласования и синтаксических ошибок

Когад-то в России и правда жило
беспечальное юное поколение,
которое улыбнулось к лету, морю и
солнцу – и выбрали «Пепси».

Щас уже турдно установить, почему
это произошло...



Анализ и коррекция текста

Исправление опечаток, ошибок согласования и синтаксических ошибок

Когад-то в России и правда жило
беспечальное юное поколение,
которое улыбнулось к лету, морю и
солнцу – и выбрали «Пепси».



Щас уже турдно установить, почему
это произошло...

Анализ и коррекция текста

Исправление опечаток, ошибок согласования и синтаксических ошибок

Когда-то в России и правда жило
беспечальное юное поколение,
которое улыбнулось лету, морю и
солнцу – и выбрало «Пепси».



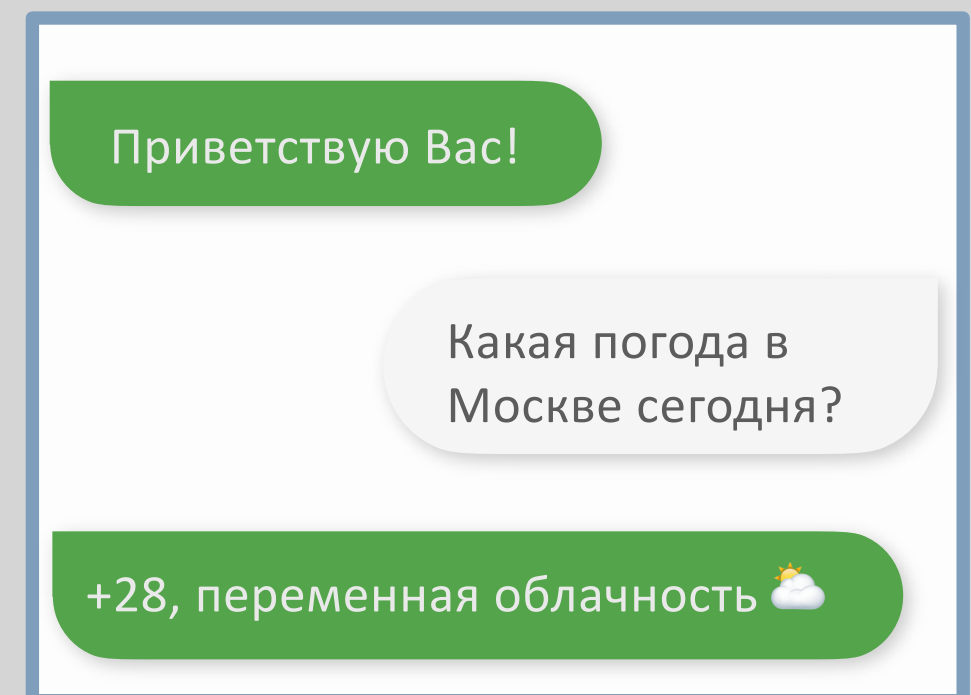
Сейчас уже трудно установить,
почему это произошло...

© В.О. Пелевин, “Generation П”

Ведение диалога

Диалоговые помощники, виртуальные ассистенты, чат-боты

- Система анализирует входных сообщений
- Система синтезирует выходных сообщений или выполняет нужное действие



Поиск ответа на вопрос

Важная часть поисковых и диалоговых систем

- Фактологические вопросы
- Вопросы по тексту
- Вопросы на здравый смысл

Кто написал “Старик и Море”?

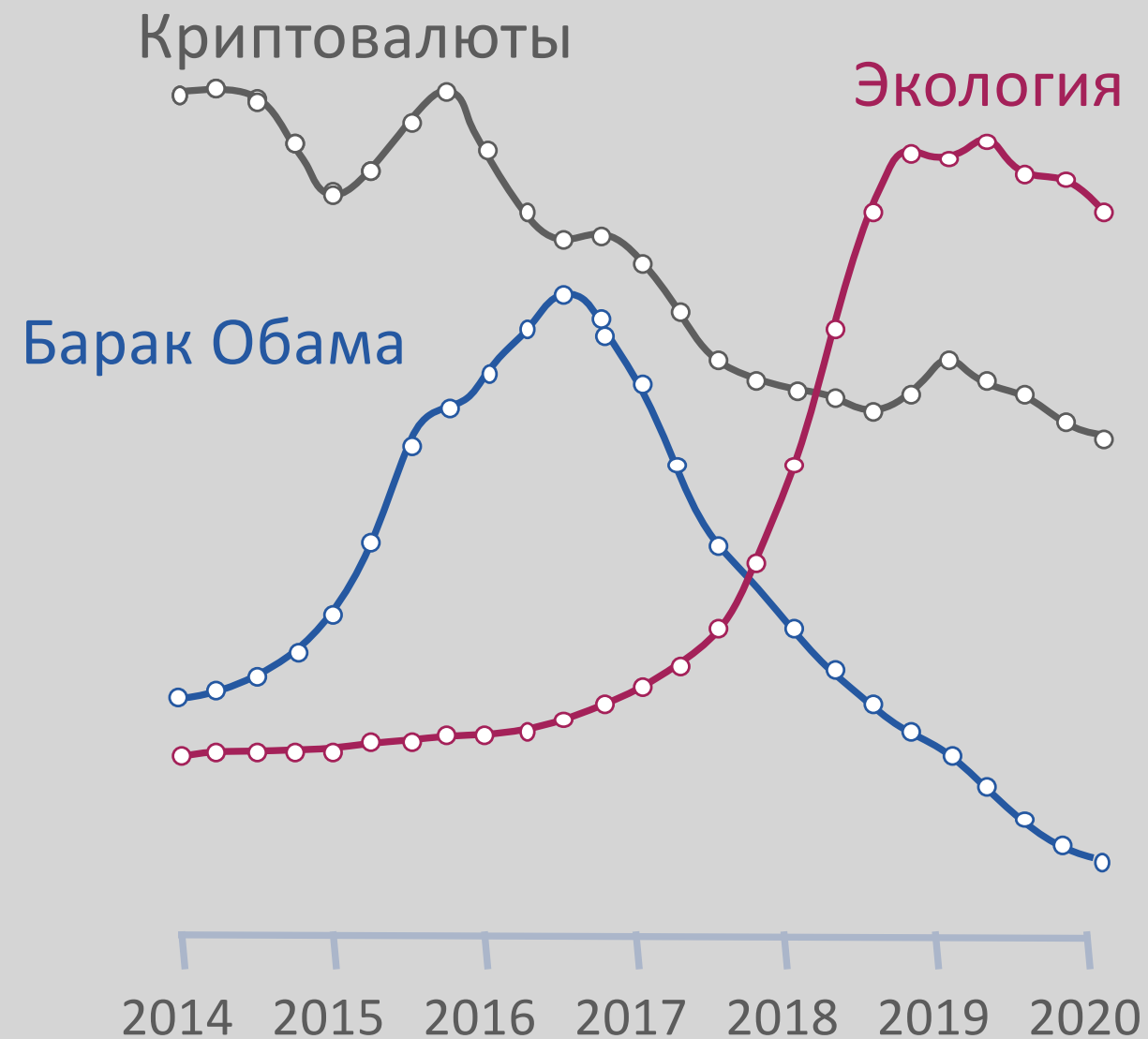
Эрнест Хемингуэй

Столица Соединённого
Королевства?

Лондон

Тренд-аналитика

Тренды в социальных сетях, бренд аналитика



Суммаризация

Получить краткое изложение длинного текста

Вот растение, которому в наш суматошный век истрёпанных нервов, изнурительных бессонниц и сдвинутой с места психики надо бы поставить красивый памятник: валериана, подобно матери, успокоит и усыпит, вернет так необходимое всем нам душевное равновесие.



© В. Солоухин

Суммаризация

Получить краткое изложение длинного текста

Валериане следовало бы поставить памятник: она успокаивает, помогает заснуть, возвращает душевное равновесие.



Связь с лингвистикой

Лингвистика

Кто такие лингвисты?

- **Лингвистика** — наука, изучающая языки
- **Лингвист** — прежде всего специалист, исследующий и описывающий языки, и иностранные, и свой родной, и даже язык вообще, универсальный механизм, обеспечивающий наше общение.

Примеры лингвистических вопросов

- Генетическое родство языков (родственен ли один язык другому)
- Бывают ли языки без гласных? Без падежей? Без глаголов?
- Когда в русском языке прилагательное идет после существительного (*мальчик красивый*, а не *красивый мальчик*)?
- Законы постановки ударений (директорá или дирéкторы? Неводы или невода?)
- Проблема границы язык/диалект
- Документация исчезающих языков

Лингвистика

... и смежные дисциплины

- **Психолингвистика** — наука на стыке нейropsихологии и лингвистики: как язык хранится в мозгу? Как дети усваивают язык и учатся говорить? Как хранятся несколько языков в мозгу у билингвов?
- **Социолингвистика** — наука на стыке социологии и лингвистики: социально обусловленное переключение языков и, шире, кодов; функционирование языка в обществе

Немного истории

Информационный поиск

- С 1940-х: информационный поиск
- Поиск фильмов, поиск подходящих книг в библиотеке
- NLP помогает лингвистике

Немного истории

Машинный перевод

- 7.01.1954: Джорджтаунский эксперимент
- Перевод около 60 предложений с русского на английский
- В основе системы лежал небольшой словарь и несколько правил
- Лингвистика помогает NLP

Немного истории

Чат-боты

- 1966: Первый чат-бот ELIZA, “психотерапевт”
- Синтаксический разбор предложения (подлежащее, сказуемое)
- Выделение ключевых слов (по списку)
- Выбор нужного шаблона ответа
- Лингвистика помогает NLP

П: У меня **болит голова**.

Д: Почему вы говорите, что у вас **болит голова**?

П: Мой **отец** меня **ненавидит**.

Д: Кто еще из **семьи** вас **ненавидит**?

Немного истории

Корпусная лингвистика

- С 1960-х: корпуса текстов
- Автоматическая разметка текстов по разным признакам, в первую очередь грамматическим
- Корпуса помогают лингвистам отвечать на вопросы типа “а когда в русском языке прилагательное может идти после существительных”
- Дисциплина “корпусная лингвистика”
- С 29.04.2004 открыт Национальный корпус русского языка
- Максимально лингвистическая часть NLP помогает лингвистике

А как сейчас?

Основной метод – deep learning

- Большинство задач, связанных с обработкой текста, решают с помощью нейронных сетей, черных ящиков
- Большая часть шагов работы с текстом (предобработка, токенизация, обучение модели) не основаны на правилах
- Лингвистика как наука не получает от черных ящиков никакой полезной информации
- Черные ящики не пользуются достижениями лингвистики
- Взаимосвязи больше нет?

А как сейчас?

Основной метод – deep learning

- Появляется ряд исследований, в которых анализируется, какую информацию о языке знает та или иная модель
- Возможно, со временем лингвисты научатся извлекать из моделей NLP новые знания и о языке
- Пока что эти две дисциплины почти не пересекаются, но снова двигаются навстречу друг другу.

Исследования

Почему важно следить за исследованиями?

Хорошие зарплаты в R&D, академический туризм ... :)

- Автоматическая обработка текстов – новая и быстро развивающаяся область
- Новые методы и модели открывают возможности для приложений, позволяют сократить время на разработку и повысить качество существующих решений
- Исследовательские разработки, как правило, публикуются в открытом доступе и могут быть использованы в коммерческих проектах

Направления исследований

- Как представить слово и текст в понятном компьютеру виде?
- Как научить компьютер понимать разные языки?
- Как научить компьютер понимать не только текст?
- Как решать одновременно несколько задач?
- Как извлечь из текстов структуру?
- Как используют методы автоматической обработки текстов в медицине?
- Как используют методы автоматической обработки текстов для анализа пользовательских текстов

Модели представления слова и текста

Первое поколение: векторное представление текста

Мешок слов – это простая модель представления текста, предложенная в середине XX-ого века. В ней не учитывается ни порядок слов, ни наличие семантических связей между ними.



абрикос

модель

слово

ящур

Модели представления слова и текста

Второе поколение: векторное представление слова

“Матрица слово-текст строится по большому корпусу”

“матрица” “слово” “документ”

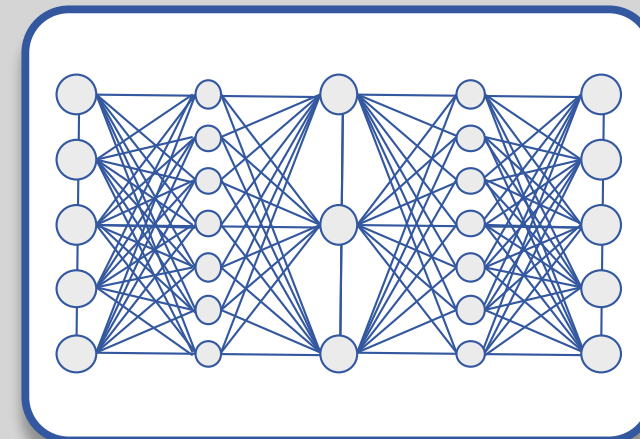
“большой” “корпус”

$\begin{bmatrix} 123 \\ 456 \\ 12 \\ \dots \\ 89 \end{bmatrix}$	$\begin{bmatrix} 23 \\ 372 \\ 8 \\ \dots \\ 83 \end{bmatrix}$	$\begin{bmatrix} 16 \\ 124 \\ 76 \\ \dots \\ 29 \end{bmatrix}$	$\begin{bmatrix} 2 \\ 12 \\ 299 \\ \dots \\ 65 \end{bmatrix}$	$\begin{bmatrix} 177 \\ 6 \\ 504 \\ \dots \\ 304 \end{bmatrix}$
---	---	--	---	---

Модели представления слова и текста

Третье поколение: языковые модели

В середине 2000-ных годов использование нейросетевых технологий стало де-факто индустриальным и академическим стандартом.

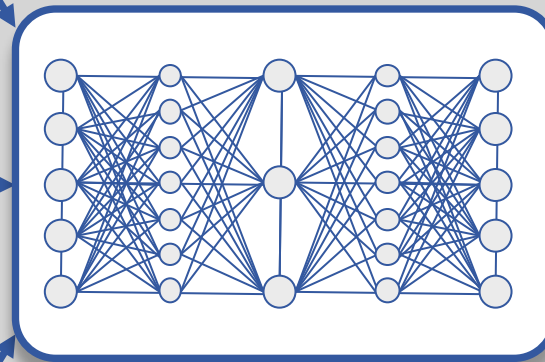

$$\begin{bmatrix} 193 & \dots & 342 \\ \vdots & \ddots & \vdots \\ 858 & \dots & 276 \end{bmatrix}$$

Мультиязычные модели

In der Computerlinguistik (CL) oder linguistischen Datenverarbeitung wird untersucht, wie natürliche Sprache in Form von Text- oder Sprachdaten mit Hilfe des Computers algorithmisch verarbeitet werden kann.

Компьютерная лингвистика — научное направление в области математического и компьютерного моделирования интеллектуальных процессов у человека и животных при создании систем искусственного интеллекта, которое ставит своей целью использование математических моделей для описания естественных языков.

Computational linguistics is an interdisciplinary field concerned with the statistical or rule-based modeling of natural language from a computational perspective, as well as the study of appropriate computational approaches to linguistic questions.



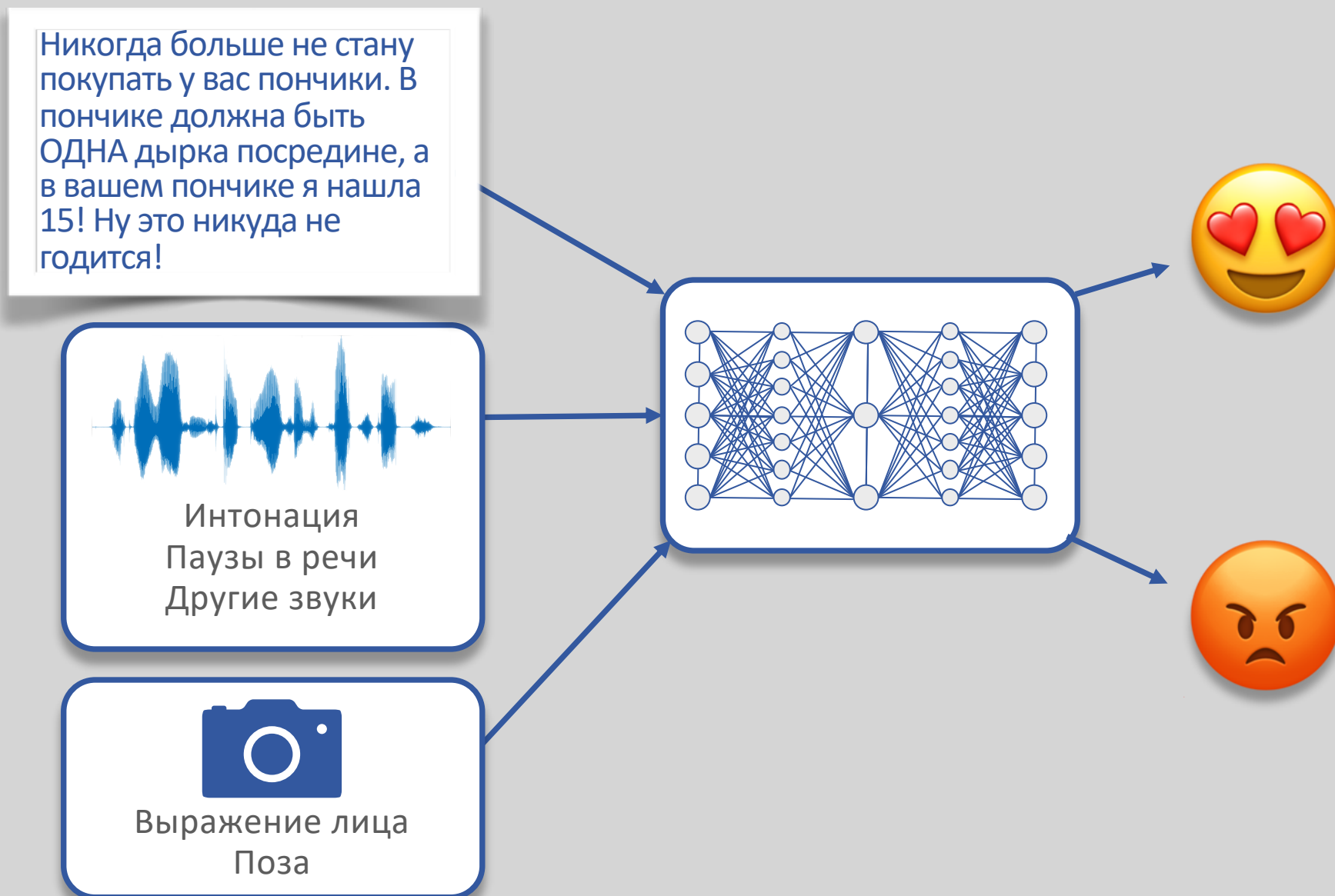
язык
language
Sprache

слово
word
Wort

модель
Modell
model

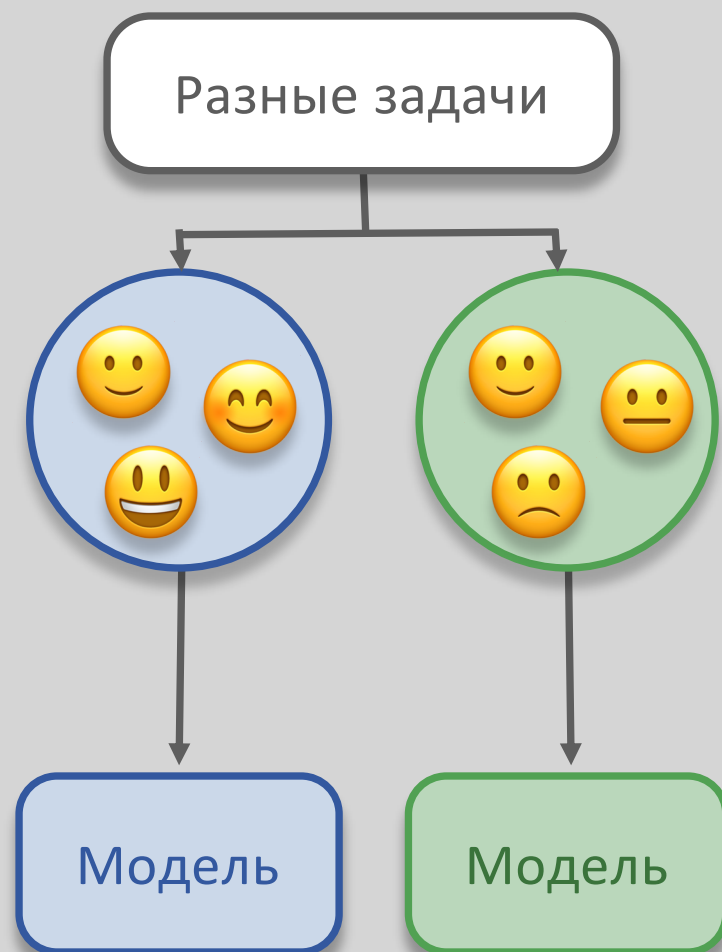
вычисления
Berechnungen
computation

Мультимодальные модели

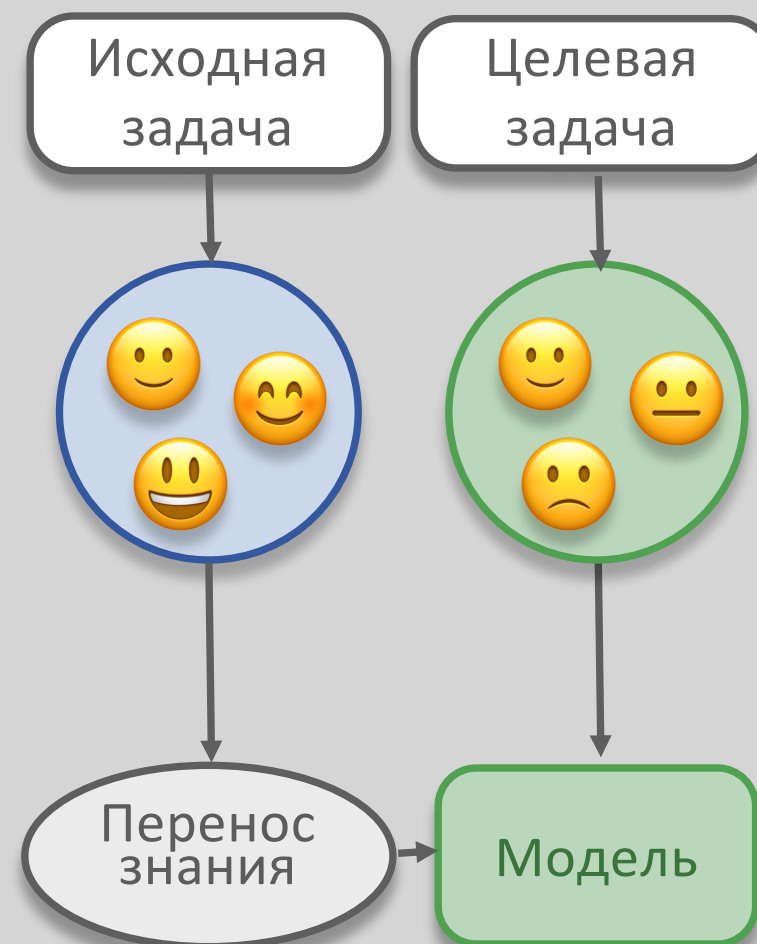


Перенос обучения

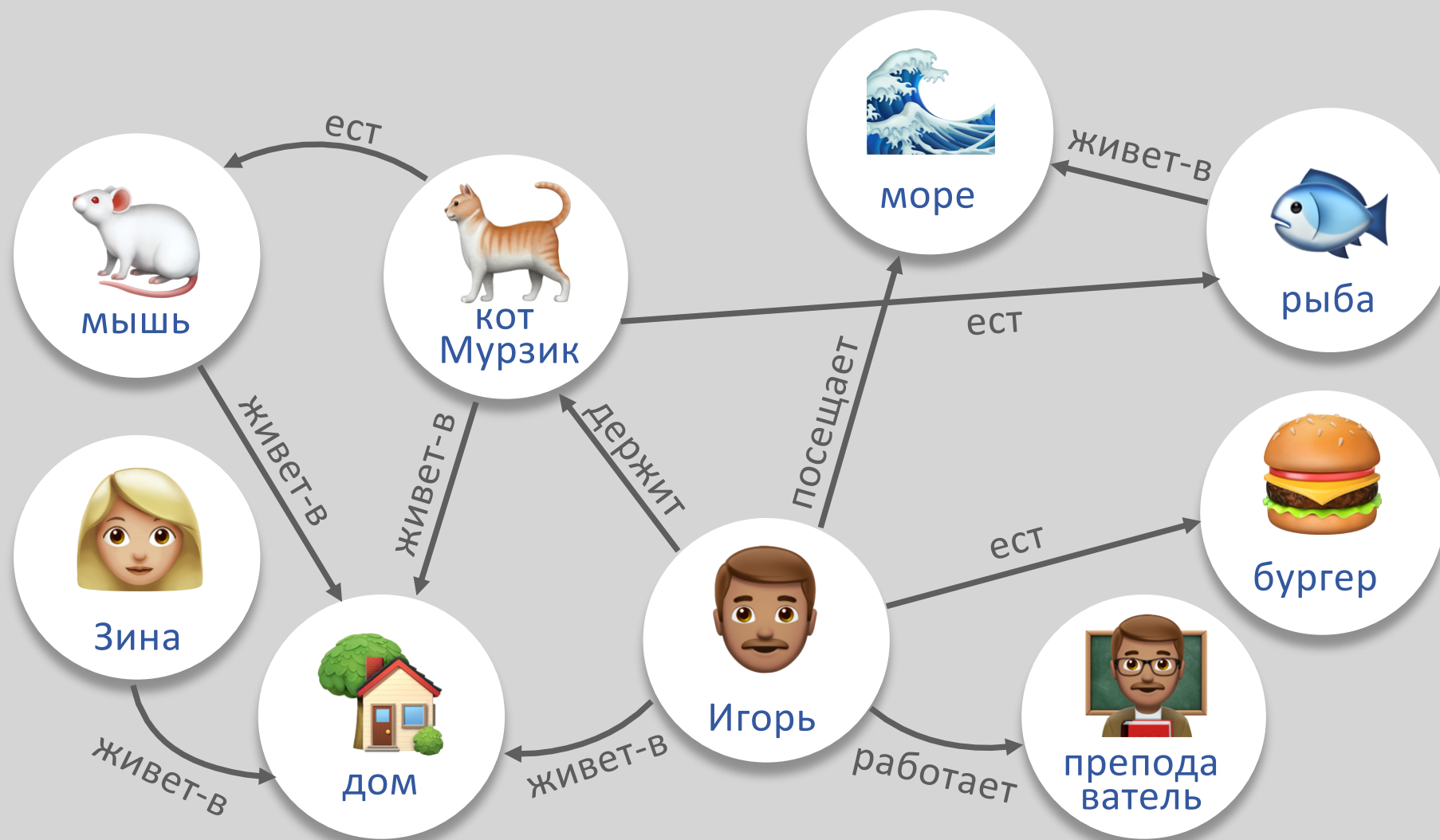
Традиционный подход к
машинному обучению



Подход на основе
переноса обучения



Графы знаний



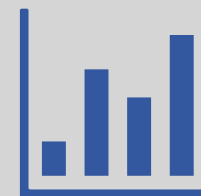
Анализ текстов в медицине



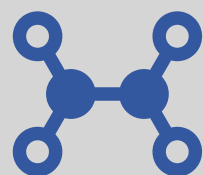
Автоматизация
записи на прием



Анализ историй
болезни



Автоматические
отчеты о больнице



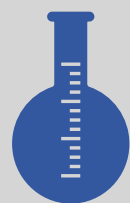
Анализ научных
статей



Анализ отзывов
на лекарства



Помощь в определении
диагноза



Поиск клинических
испытаний

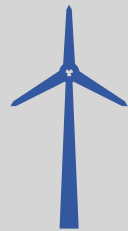


Телемедицина



Ответ на частые
вопросы

Анализ социальных сетей и социальных медиа



Как заботятся об
экологии?



Как выбирают
смартфоны?



Что покупают для
детей?



Где отдыхают?



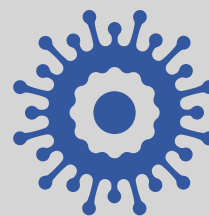
Какие бывают
пищевые привычки?



Как относятся к
политике?



Что дарят?



Как передаются
вирусы?



Как возникают
ложные новости?



Компьютерный юмор



Предобработка

Предобработка текстовых данных

Основные этапы

- Токенизация
- Сегментация предложений
- Удаление пунктуации
- Удаление стоп-слов
- Фильтрация по длине, частоте, регулярному выражению
- Лемматизация – приведение к нормальной форме ()
- Стемминг – приведение к псевдооснове ()

Предобработка текстовых данных

Основные этапы

- Токенизация
 - По пробельному символу
 - Регулярные выражения
- Сегментация предложений
 - Регулярные выражения
 - Классификаторы для сложных случаев

Предобработка текстовых данных

Основные этапы

- Удаление пунктуации
 - Смайлики нужны для классификации по тональности
 - Дефисы и тире часто перепутаны
 - Регулярные выражения
- Регистр
 - В некоторых задачах важен регистр
 - А в некоторых – создает избыточное пространство признаков

Предобработка текстовых данных

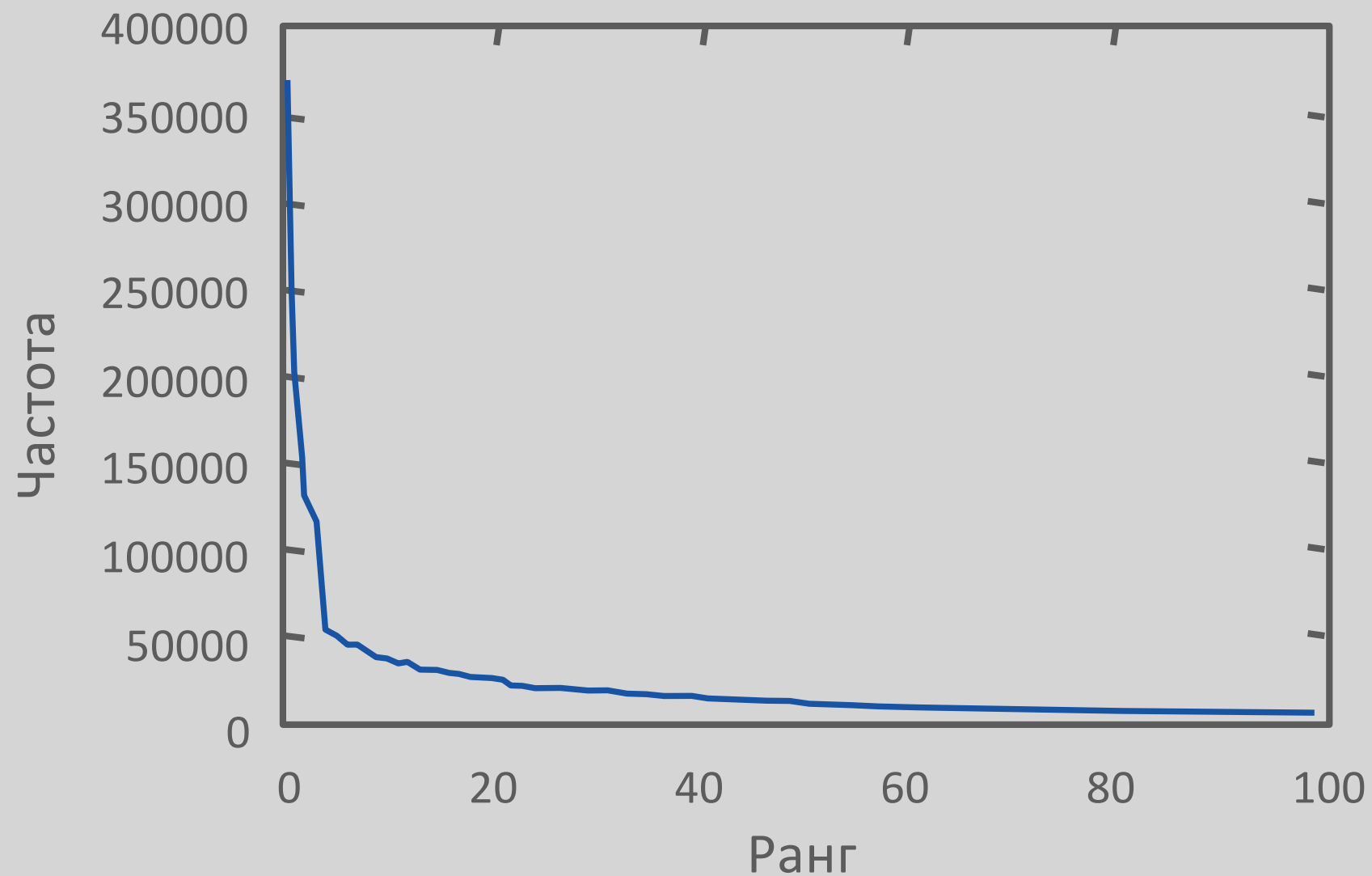
Основные этапы

- Удаление стоп-слов
 - Союзы
 - Предлоги
 - Местоимения
 - Вспомогательные и модальные глаголы
- Удаление слишком частых и самых редких слов

Предобработка

Основные этапы

Закон Ципфа



Предобработка

Лемматизация – приведение к нормальной форме

Туристам
очень
понравилась
прогулка по
мосту



Турист очень
понравиться
прогулка по
мост

Предобработка

Стемминг – приведение слова к псевдооснове

Туристам
очень
понравилась
прогулка по
мосту



Турист очень
понрав
прогулк по
мост

Векторизация

Векторное представление документа

Модель “мешка слов”

Номер текста	вхождений слова «абрикос»	...	вхождений слова «ямал»
1	0	...	23
...
N	4	...	0

Векторное представление документа

$tf - idf$ взвешивание показывает насколько важно слово t и насколько оно отличает документ d от прочих

- $d \in D$ – множество документов

- $t \in V$ – множество слов

- $tf_{t,d} = \frac{\text{count}(t, d)}{\sum_{t' \in V} \text{count}(t', d)}$

- $df_{t,d,D} = | \{ d \in D : \text{count}(t, d) > 0 \} |$

- $tf - idf(t, d, D) = tf_{t,d} \times \log \frac{|D|}{df_{t,d,D} + 1}$

Векторное представление документа

Модель “мешка слов”

- Проблемы
 - Не учитывается порядок слов
 - Не учитываются синонимы

Векторное представление документа

Выделение коллокаций

- Статистические меры, помогающие учесть зависимость между словами
- N_{t_1, t_2} – число документов, содержащих оба слова
- N_{t_1} – число документов, содержащих одно слово

$$\text{PMI}(t_1, t_2) = \log \frac{N_{t_1, t_2}}{N_{t_1} N_{t_2}}$$

Векторное представление документа

Выделение коллокаций

PMI	w^1	w^2
4.4862	Стивен	Хокинг
4.4862	Чарли	Чаплин
4.4860	Альбрехт	Дюрер
4.4860	светлое	пиво
4.4860	холодное	оружие
2.3764	корзина	интерес
2.3541	гастроном	произведение
1.4875	базар	чай
1.2456	кресло	папироса
0.8752	задание	колодец

Заключение

Основные выводы

- Для текстов можно строить признаковые описания
- Мы хотим учитывать в признаках как можно больше информации, содержащейся в тексте
- Стандартными признаковыми описаниями являются «мешок слов» и векторы $tf - idf$
- N-граммные признаки обычно позволяют обучать более качественные модели, чем униграммные