ЛЕКЦИЯ 9

МЕТОДЫ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Основные понятия • определения • схемы • таблицы • графики • примеры

Преподаватель: проф.Ахмедиярова А.Т.

Структура и основные разделы

01 Понятие обработки естественного языка

02 Особенности естественного языка

03 Общая схема обработки текста

04 Очистка текстовых данных

05 Токенизация

06 Нормализация текста

07 Стоп-слова

08 Мешок слов — Bag of Words

09 TF-IDF

10 Векторные представления слов

11 Контекстные языковые представления

12 Классификация текстов

Понятие обработки естественного языка

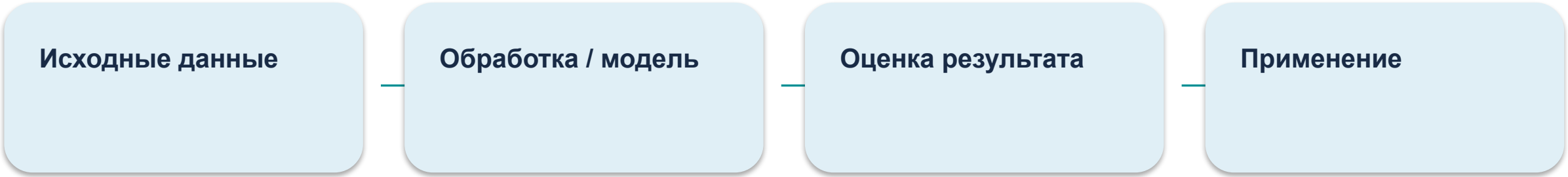
Обработка естественного языка (Natural Language Processing, NLP) — направление искусственного интеллекта, связанное с компьютерной обработкой, анализом, пониманием и генерацией человеческого языка. Текст является неструктурированным типом данных, поэтому перед применением алгоритмов машинного обучения его необходимо преобразовать в форму, пригодную для вычислительной обработки.

Ключевая идея: Обработка естественного языка (Natural Language Processing, NLP) — направление искусственного интеллекта, связанное с компьют

Понятие обработки естественного языка — продолжение

К задачам NLP относятся классификация текстов, анализ тональности, извлечение информации, поиск именованных сущностей, тематический анализ, автоматическое реферирование, семантический поиск, ответы на вопросы и машинный перевод.

Схема: Понятие обработки естественного языка



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Особенности естественного языка

Естественный язык сложен для автоматической обработки, поскольку одно слово может иметь несколько значений, смысл зависит от контекста, а одинаковая мысль может выражаться разными словами и грамматическими конструкциями. Дополнительные трудности создают опечатки, сокращения, разговорные формы, специальные термины и смешение языков.

Ключевая идея: Естественный язык сложен для автоматической обработки, поскольку одно слово может иметь несколько значений, смысл зависит от

Особенности естественного языка — продолжение

Например, слово «ключ» может обозначать инструмент, источник воды или криптографический ключ. Определить нужное значение можно только с учётом окружающего контекста. Поэтому современные NLP-системы стремятся учитывать не только отдельные слова, но и связи между ними.

Сравнительная таблица: Особенности естественного языка

Особенности естественного языка	Естественный язык сложен для автоматической обработки, поскольку одно слово может иметь несколько значений, смысл з
Общая схема обработки текста	Например, слово «ключ» может обозначать инструмент, источник воды или криптографический ключ. Определить нужное зна

Общая схема обработки текста

Типичный NLP-конвейер включает получение текстовых данных, очистку, токенизацию, нормализацию, преобразование текста в числовое представление, обучение или применение модели и оценку результата. Конкретный набор этапов зависит от задачи и используемой модели.

Очистка текстовых данных

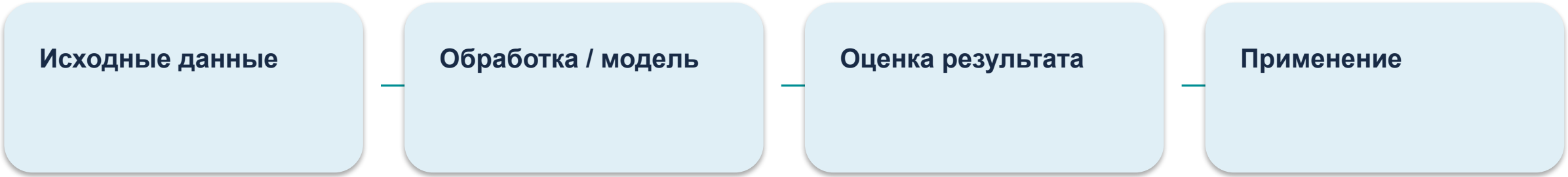
Очистка текста устраняет элементы, которые не несут полезной информации для конкретной задачи или создают технический шум. Возможные операции включают удаление лишних пробелов, служебных символов, HTML-тегов, повторяющихся знаков, приведение регистра и обработку ссылок. Однако очистка не должна выполняться механически.

Ключевая идея: Очистка текста устраняет элементы, которые не несут полезной информации для конкретной задачи или создают технический шум. Во

Очистка текстовых данных — продолжение

Например, знаки пунктуации могут быть важны для анализа тональности, а регистр — для распознавания именованных сущностей. Поэтому набор операций определяется целью исследования.

Схема: Очистка текстовых данных



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Токенизация

Токенизация — разбиение текста на отдельные единицы, называемые токенами. Токенами могут быть слова, части слов, символы или предложения. Это один из базовых этапов обработки текста. Пример: предложение «Модель обнаружила подозрительное соединение» при простой токенизации по словам может быть представлено как [«Модель», «обнаружила», «подозрительное», «соединение»].

Ключевая идея: Токенизация — разбиение текста на отдельные единицы, называемые токенами. Токенами могут быть слова, части слов, символы или

Нормализация текста

После токенизации текст может быть нормализован. Нормализация уменьшает количество различных форм, которые модель должна учитывать. К распространённым операциям относятся приведение к нижнему регистру, удаление или специальная обработка пунктуации, лемматизация и стемминг. Лемматизация приводит слово к словарной форме с учётом морфологии.

Ключевая идея: После токенизации текст может быть нормализован. Нормализация уменьшает количество различных форм, которые модель должна учит

Нормализация текста — продолжение

Например, «документы», «документа» и «документами» могут быть приведены к лемме «документ». Стемминг обычно механически сокращает слово до основы и может давать форму, не являющуюся полноценным словом.

Сравнительная таблица: Нормализация текста

Нормализация текста	После токенизации текст может быть нормализован. Нормализация уменьшает количество различных форм, которые модель д
Стоп-слова	Лемматизация приводит слово к словарной форме с учётом морфологии. Например, «документы», «документа» и «документам

Стоп-слова

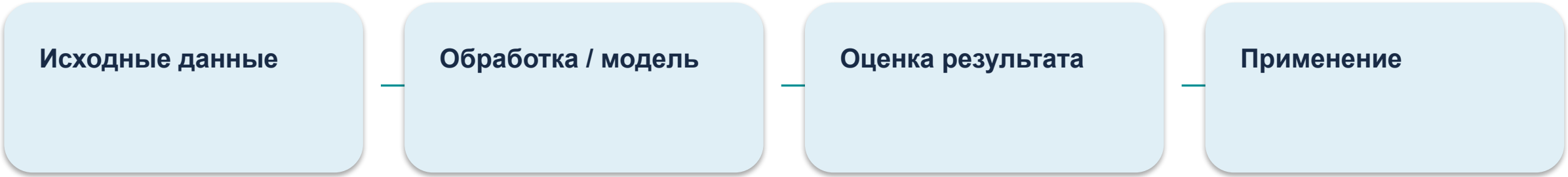
Стоп-слова — часто встречающиеся слова, которые в некоторых задачах имеют небольшую различительную ценность. К ним могут относиться союзы, предлоги и некоторые служебные слова. В классических моделях их иногда удаляют для уменьшения размерности.

Ключевая идея: Стоп-слова — часто встречающиеся слова, которые в некоторых задачах имеют небольшую различительную ценность. К ним могут отно

Стоп-слова — продолжение

В современных контекстных моделях стоп-слова часто сохраняют, поскольку они участвуют в формировании смысла и грамматической структуры предложения. Следовательно, удаление стоп-слов не является обязательным правилом.

Схема: Стоп-слова



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Мешок слов — Bag of Words

Bag of Words представляет документ через частоты слов из заранее сформированного словаря. Порядок слов обычно не учитывается. Если словарь содержит m терминов, каждый документ превращается в вектор длины m. Пример. Для словаря [«атака», «сервер», «доступ»] документ «атака на сервер» можно представить вектором [1, 1, 0].

Ключевая идея: Bag of Words представляет документ через частоты слов из заранее сформированного словаря. Порядок слов обычно не учитывается.

Мешок слов — Bag of Words — продолжение

Такой подход прост и интерпретируем, но плохо учитывает контекст и семантику.

TF-IDF

TF-IDF — способ числового представления текста, который учитывает частоту термина в документе и его распространённость во всей коллекции. Термин получает больший вес, если он часто встречается в конкретном документе, но сравнительно редко встречается в других документах.

Ключевая идея: TF-IDF — способ числового представления текста, который учитывает частоту термина в документе и его распространённость во все

TF-IDF — продолжение

Упрощённо TF-IDF можно представить как произведение двух величин: TF-IDF(t, d) = TF(t, d) × IDF(t) TF показывает частоту термина t в документе d, а IDF уменьшает вес слов, встречающихся во многих документах. В результате более специфичные для документа термины получают больший вес.

TF-IDF — продолжение

Пример TF-IDF в Python from sklearn.feature_extraction.text import TfidfVectorizer texts = ['подозрительная попытка входа в систему', 'успешный вход пользователя в систему', 'обнаружена подозрительная сетевая активность'] vectorizer = TfidfVectorizer() X = vectorizer.fit_transform(texts) print(vectorizer.get_feature_names_out()) print(X.toarray()) В результате каждый текст преобразуется в числовой вектор.

TF-IDF — продолжение

Полученную матрицу можно передать алгоритму классификации, кластеризации или поиска сходства.

Векторные представления слов

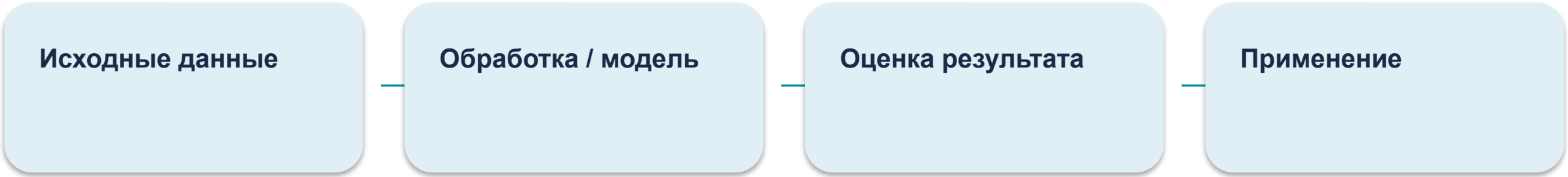
Ограничение Bag of Words и TF-IDF состоит в том, что смысловые отношения между словами напрямую не моделируются. Векторные представления слов (word embeddings) отображают слова в плотное числовое пространство, где семантически близкие слова могут иметь близкие векторы. В загруженных материалах среди методов NLP указаны Word2Vec и GloVe.

Ключевая идея: Ограничение Bag of Words и TF-IDF состоит в том, что смысловые отношения между словами напрямую не моделируются. Векторные пр

Векторные представления слов — продолжение

Такие подходы позволяют перейти от разреженных частотных векторов к плотным представлениям, отражающим статистические отношения между словами.

Схема: Векторные представления слов



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Сравнительная таблица: Векторные представления слов

Векторные представления слов	Ограничение Bag of Words и TF-IDF состоит в том, что смысловые отношения между словами напрямую не моделируются. Ве
Контекстные языковые представления	В загруженных материалах среди методов NLP указаны Word2Vec и GloVe. Такие подходы позволяют перейти от разреженных

Контекстные языковые представления

Современные нейросетевые методы формируют представление слова с учётом окружающего текста. Поэтому одно и то же слово в разных предложениях может иметь разные векторы. В загруженных материалах среди NLP-подходов указаны BERT и GPT-модели, способные лучше учитывать контекст и семантические связи, но предъявляющие более высокие требования к вычислительным ресурсам.

Ключевая идея: Современные нейросетевые методы формируют представление слова с учётом окружающего текста. Поэтому одно и то же слово в разны

Контекстные языковые представления — продолжение

Например, значение слова «ключ» в выражениях «ключ шифрования» и «ключ от двери» различается. Контекстная модель стремится сформировать разные представления этих употреблений.

Классификация текстов

Классификация текста — задача присвоения документу одной или нескольких категорий. Классический подход включает очистку текста, TF-IDF-векторизацию и обучение классификатора, например логистической регрессии, SVM или Naive Bayes.

Ключевая идея: Классификация текста — задача присвоения документу одной или нескольких категорий. Классический подход включает очистку текст

Классификация текстов — продолжение

Примеры: классификация обращений пользователей по темам; определение фишинговых сообщений; распределение документов по категориям; классификация событий информационной безопасности по типу инцидента.

Классификация текстов — продолжение

Пример классификации текста from sklearn.model_selection import train_test_split from sklearn.feature_extraction.text import TfidfVectorizer from sklearn.linear_model import LogisticRegression from sklearn.metrics import classification_report X_train, X_test, y_train, y_test = train_test_split(texts, labels, test_size=0.2, random_state=42, stratify=labels) vectorizer = TfidfVectorizer(ngram_range=(1, 2)) X_train_vec = vectorizer.fit_transform(X_train) X_test_vec = vectorizer.transform(X_test) model = LogisticRegression(max_iter=1000) model.fit(X_train_vec, y_train) y_pred = model.predict(X_test_vec) print(classification_report(y_test, y_pred))

N-граммы

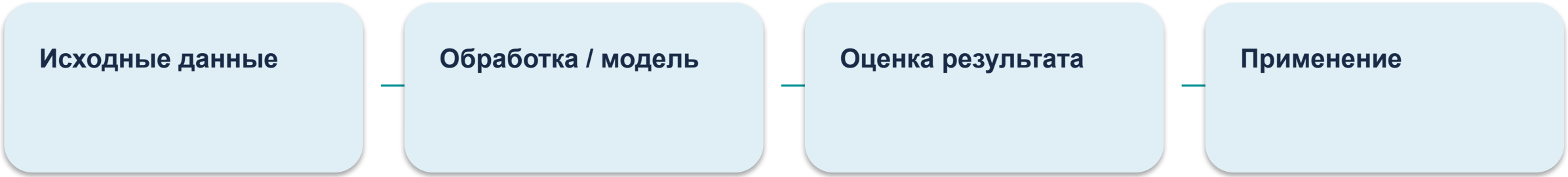
N-грамма — последовательность из n соседних токенов. Unigram содержит один токен, bigram — два, trigram — три. N-граммы позволяют частично учитывать локальный порядок слов в классических векторных моделях. Например, слова «не» и «безопасно» по отдельности могут не передавать полный смысл, а bigram «не безопасно» содержит более информативную комбинацию.

Ключевая идея: N-грамма — последовательность из n соседних токенов. Unigram содержит один токен, bigram — два, trigram — три. N-граммы позво

N-граммы — продолжение

Однако увеличение числа n резко увеличивает размер словаря.

Схема: N-граммы



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Анализ тональности

Анализ тональности (sentiment analysis) определяет эмоциональную или оценочную окраску текста. Базовая постановка включает положительный, отрицательный и нейтральный классы, хотя в прикладных системах возможны более сложные шкалы.

Ключевая идея: Анализ тональности (sentiment analysis) определяет эмоциональную или оценочную окраску текста. Базовая постановка включает по

Анализ тональности — продолжение

Например, отзывы пользователей о цифровом сервисе можно автоматически разделить на положительные и отрицательные, а затем анализировать наиболее частые причины недовольства. Для юридических и технических документов термин «тональность» может быть менее естественным, поэтому постановка задачи должна соответствовать типу текста.

Сравнительная таблица: Анализ тональности

Анализ тональности	Анализ тональности (sentiment analysis) определяет эмоциональную или оценочную окраску текста. Базовая постановка в
Извлечение именованных сущностей NER	Например, отзывы пользователей о цифровом сервисе можно автоматически разделить на положительные и отрицательные, а

Извлечение именованных сущностей NER

Named Entity Recognition (NER) — выделение в тексте сущностей определённых типов. К типичным категориям относятся люди, организации, географические объекты, даты и другие предметно-ориентированные сущности. В юридическом документе NER может использоваться для поиска участников, организаций, дат и других реквизитов.

Ключевая идея: Named Entity Recognition (NER) — выделение в тексте сущностей определённых типов. К типичным категориям относятся люди, орган

Извлечение именованных сущностей NER — продолжение

В журналах информационной безопасности предметно-ориентированная система может извлекать IP-адреса, домены, имена файлов и идентификаторы, если такие категории предусмотрены моделью или правилами.

Тематическое моделирование

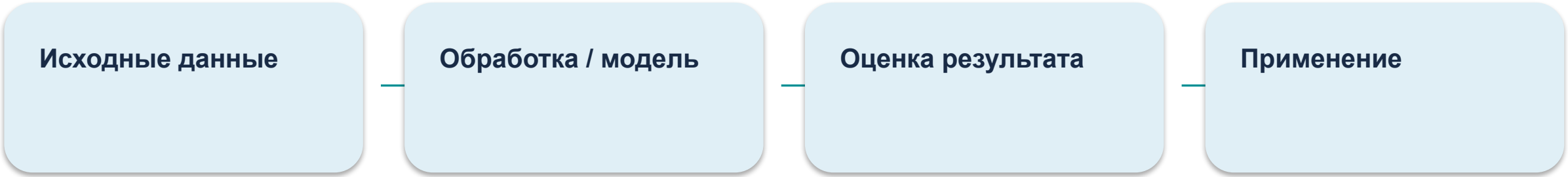
Тематическое моделирование предназначено для выявления скрытых тематических структур в коллекции документов. В загруженных материалах среди методов анализа текста указано LDA. Модель рассматривает документы как сочетания тем, а темы — как распределения слов.

Ключевая идея: Тематическое моделирование предназначено для выявления скрытых тематических структур в коллекции документов. В загруженных ма

Тематическое моделирование — продолжение

Такой подход может применяться для первичного исследования больших коллекций обращений или документов, когда заранее неизвестно, какие тематические группы присутствуют в корпусе.

Схема: Тематическое моделирование



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Семантический поиск

Семантический поиск стремится находить документы не только по совпадению слов, но и по смысловой близости. Для этого запрос и документы преобразуются в векторные представления, после чего вычисляется их сходство. Например, запрос «несанкционированный доступ к аккаунту» может быть смыслово близок документу, содержащему выражение «взлом учётной записи», даже если точные слова не совпадают.

Ключевая идея: Семантический поиск стремится находить документы не только по совпадению слов, но и по смысловой близости. Для этого запрос и

Семантический поиск — продолжение

Это важное отличие от простого лексического поиска.

Косинусное сходство

Для сравнения текстовых векторов часто используется косинусное сходство. Оно оценивает угол между двумя векторами и меньше зависит от их абсолютной длины: cos(x, y) = (x · y) / (|| x|| ||y||) Чем ближе значение к 1 для неотрицательных текстовых представлений, тем более сходным считается направление векторов. Косинусное сходство широко применяется к TF-IDF и embedding-представлениям.

Ключевая идея: Для сравнения текстовых векторов часто используется косинусное сходство. Оно оценивает угол между двумя векторами и меньше за

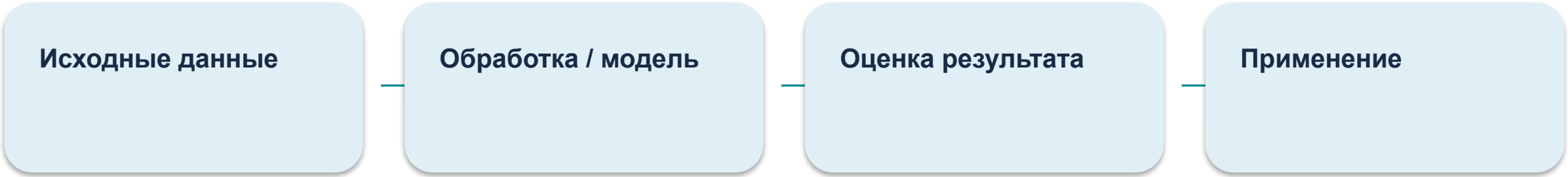
Сравнительная таблица: Косинусное сходство

Косинусное сходство	Для сравнения текстовых векторов часто используется косинусное сходство. Оно оценивает угол между двумя векторами и
Суммаризация текста	cos(x, y) = (x · y) / (||x|| ||y||)
Машинный перевод	Чем ближе значение к 1 для неотрицательных текстовых представлений, тем более сходным считается направление векторо

Суммаризация текста

Суммаризация предназначена для получения краткого представления длинного текста. Экстрактивные методы выбирают наиболее значимые фрагменты исходного документа. Абстрактивные методы формируют новое краткое изложение содержания.

Схема: Суммаризация текста



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Машинный перевод

Машинный перевод — автоматическое преобразование текста с одного естественного языка на другой. В учебном пособии машинный перевод включён в основные темы главы NLP. Современные системы перевода обычно используют нейросетевые архитектуры и обучаются на больших параллельных или многоязычных корпусах.

Ключевая идея: Машинный перевод — автоматическое преобразование текста с одного естественного языка на другой. В учебном пособии машинный пе

Машинный перевод — продолжение

Для казахского, русского и английского языков особое значение имеют качество многоязычных данных, морфологические особенности языков, терминология предметной области и корректная передача именованных сущностей.

NLP для казахского и русского языков

Многоязычная обработка требует учитывать различия языков. Казахский язык обладает развитой словоизменительной системой, поэтому одна основа может встречаться во множестве словоформ. Для русского языка также важны морфологические формы, падежи и согласование. В смешанных сообщениях дополнительно встречается переключение между языками.

Ключевая идея: Многоязычная обработка требует учитывать различия языков. Казахский язык обладает развитой словоизменительной системой, поэто

NLP для казахского и русского языков — продолжение

При разработке системы необходимо проверять качество токенизатора, нормализации и модели отдельно на каждом целевом языке, а не переносить результаты одного языка на другой без экспериментальной проверки.

NLP в информационной безопасности

NLP не заменяет технические средства обнаружения атак. Он дополняет их, позволяя анализировать текстовые источники, сообщения, отчёты и описания событий. Пример: анализ фишингового сообщения Рассмотрим задачу бинарной классификации электронных сообщений. Обучающая выборка содержит тексты и метки legitimate или phishing. После очистки и TF-IDF-векторизации обучается классификатор.

Ключевая идея: NLP не заменяет технические средства обнаружения атак. Он дополняет их, позволяя анализировать текстовые источники, сообщения

NLP в информационной безопасности — продолжение

Для оценки необходимо учитывать не только accuracy, но и precision и recall класса phishing, поскольку пропуск опасного сообщения и ложная блокировка легитимного письма имеют разную цену. Дополнительно могут использоваться признаки URL, домена, структуры письма и метаданных.

NLP в информационной безопасности — продолжение

Это пример мультимодального по смыслу решения, где текстовые признаки NLP объединяются с техническими характеристиками сообщения.

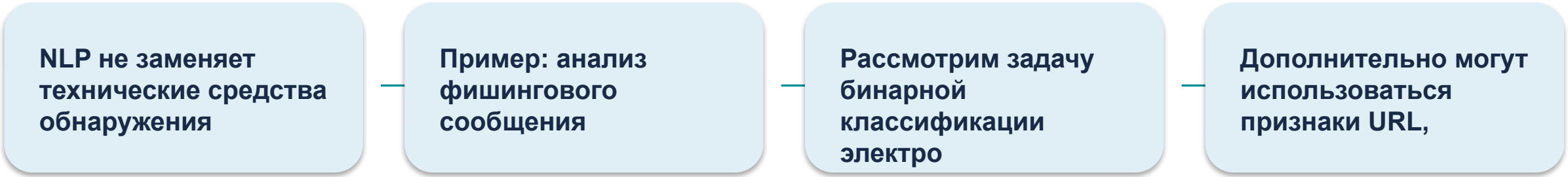
NLP в информационной безопасности — продолжение

Пример: анализ юридических документов Для большого корпуса юридических документов NLP может применяться последовательно: токенизация и очистка текста; TF-IDF или контекстные представления; извлечение сущностей; классификация документов; тематическое моделирование; суммаризация; семантический поиск.

NLP в информационной безопасности — продолжение

В загруженных проектных материалах для анализа судебных данных перечислены TF-IDF, Word2Vec, GloVe, BERT и GPT-модели, а среди задач — NER, анализ тональности, LDA, суммаризация и семантический поиск.

Схема: NLP в информационной безопасности



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Сравнительная таблица: NLP в информационной безопасности

NLP в информационной безопасности	NLP не заменяет технические средства обнаружения атак. Он дополняет их, позволяя анализировать текстовые источники,
Оценка качества NLP-моделей	Пример: анализ фишингового сообщения
Разделение данных и утечка информации	Рассмотрим задачу бинарной классификации электронных сообщений. Обучающая выборка содержит тексты и метки legitimat
Классические методы и нейросетевой NLP	Дополнительно могут использоваться признаки URL, домена, структуры письма и метаданных. Это пример мультимодального

Оценка качества NLP-моделей

Метрика зависит от задачи. Для классификации применяются accuracy, precision, recall и F1-score. Для NER также часто рассчитывают precision, recall и F1 по сущностям. Для машинного перевода и суммаризации используются специальные автоматические метрики, однако их необходимо дополнять содержательной оценкой качества текста.

Разделение данных и утечка информации

При обучении NLP-модели необходимо разделять данные на обучающую, валидационную и тестовую части. Особое внимание следует уделять утечке данных. Например, почти одинаковые версии одного документа не должны одновременно попадать в train и test, если это делает тест искусственно простым.

Ключевая идея: При обучении NLP-модели необходимо разделять данные на обучающую, валидационную и тестовую части. Особое внимание следует уде

Разделение данных и утечка информации — продолжение

Векторизатор TF-IDF также должен обучаться только на обучающей части, а к test применяться уже настроенное преобразование. Иначе информация о тестовой коллекции частично попадает в процесс подготовки признаков.

Типичные ошибки при работе с текстом

Удалять пунктуацию, регистр или стоп-слова без проверки, нужны ли они для конкретной задачи. Обучать TF-IDF на всём наборе данных до разделения train/test. Считать TF-IDF семантическим представлением, полностью учитывающим контекст. Использовать только accuracy при сильном дисбалансе классов. Не учитывать язык, морфологию и смешение языков в многоязычном корпусе.

Ключевая идея: Удалять пунктуацию, регистр или стоп-слова без проверки, нужны ли они для конкретной задачи.

Типичные ошибки при работе с текстом — продолжение

Считать автоматически сгенерированную суммаризацию или перевод безусловно корректными. Сравнивать модели на разных тестовых выборках и делать вывод о превосходстве алгоритма. Использовать сложную нейросетевую модель без сравнения с простой базовой моделью.

Сравнительная таблица: Типичные ошибки при работе с текстом

Типичные ошибки при работе с текстом	Удалять пунктуацию, регистр или стоп-слова без проверки, нужны ли они для конкретной задачи.
Практический пример полного NLP-конвейера	Обучать TF-IDF на всём наборе данных до разделения train/test.
Основные выводы	Считать TF-IDF семантическим представлением, полностью учитывающим контекст.
Контрольные вопросы	Использовать только accuracy при сильном дисбалансе классов.

Практический пример полного NLP-конвейера

Пусть необходимо автоматически классифицировать обращения пользователей информационной системы по категориям «доступ», «сетевая проблема», «подозрительная активность» и «другое». Сначала формируется размеченный корпус. Затем тексты очищаются с сохранением значимых технических обозначений, выполняется токенизация и строится TF-IDF-представление.

Ключевая идея: Пусть необходимо автоматически классифицировать обращения пользователей информационной системы по категориям «доступ», «сетев

Практический пример полного NLP-конвейера — продолжение

На одинаковом разбиении данных сравниваются Logistic Regression и SVM. После оценки confusion matrix и F1-score по каждому классу анализируются типичные ошибки. Если система плохо распознаёт смыслово близкие формулировки с разной лексикой, можно перейти к embedding-представлениям или контекстной модели.

Практический пример полного NLP-конвейера — продолжение

Таким образом, сложность решения увеличивается только при наличии экспериментальной необходимости.

Основные выводы

NLP предназначен для компьютерной обработки, анализа, понимания и генерации естественного языка. Текст требует специальной подготовки и преобразования в числовое представление. Токенизация разделяет текст на единицы, а нормализация уменьшает вариативность форм. Bag of Words и TF-IDF являются классическими способами векторизации текста.

Ключевая идея: NLP предназначен для компьютерной обработки, анализа, понимания и генерации естественного языка.

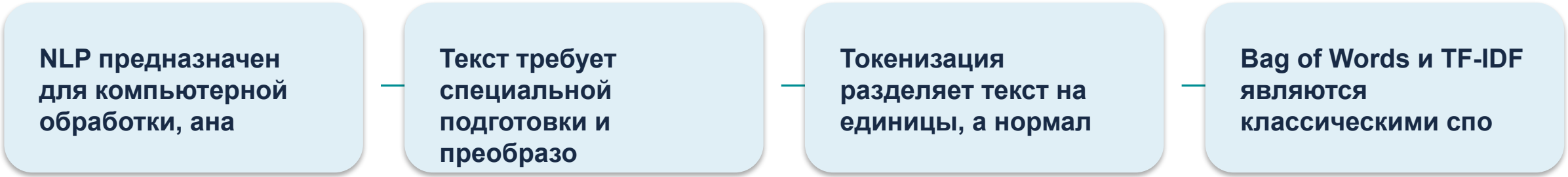
Основные выводы — продолжение

Word2Vec и GloVe формируют плотные векторные представления слов, а контекстные модели учитывают окружение слова. К основным задачам NLP относятся классификация, анализ тональности, NER, тематическое моделирование, суммаризация, семантический поиск и машинный перевод. В информационной безопасности NLP применяется к фишинговым сообщениям, журналам, отчётам об угрозах и описаниям инцидентов.

Основные выводы — продолжение

Качество NLP-системы необходимо оценивать на независимых данных и подходящими для задачи метриками.

Схема: Основные выводы



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Контрольные вопросы

1. Что понимается под обработкой естественного языка? 2. Какие основные задачи относятся к NLP? 3. Почему естественный язык сложен для автоматической обработки? 4. Что такое текстовый корпус? 5. Из каких этапов может состоять NLP-конвейер? 6. Что такое токенизация? 7. Какие виды токенизации существуют? 8. Что понимается под нормализацией текста? 9.

Ключевая идея: 1. Что понимается под обработкой естественного языка?

Контрольные вопросы — продолжение

Чем лемматизация отличается от стемминга? 10. Что такое стоп-слова и всегда ли их следует удалять? 11. Как работает Bag of Words? В чём основная идея TF-IDF? 13. Что показывает TF в TF-IDF? Какова роль IDF? Какие ограничения имеет TF-IDF? 16. Что такое word embeddings? 17. В чём идея Word2Vec и GloVe?

Список источников

Учебное пособие «Основы искусственного интеллекта». Глава 9 «Обработка естественного языка (NLP)»: токенизация, векторизация, анализ тональности, машинный перевод. Бессмертный И.А., Авдюшина А.Е., Кугаевских А.В., Королева Ю.А., Рущенко Н.Г. Системы искусственного интеллекта. Методическое пособие. Санкт-Петербург: Университет ИТМО, 2024.

Ключевая идея: Учебное пособие «Основы искусственного интеллекта». Глава 9 «Обработка естественного языка (NLP)»: токенизация, векторизация,

Список источников — продолжение

Рабочая программа дисциплины «Искусственный интеллект и машинное обучение». Направление подготовки 09.03.01 «Информатика и вычислительная техника». Москва, 2024. Сидоркина И.Г. Системы искусственного интеллекта. Учебное пособие. Москва: КНОРУС, 2015. Бондарев В.Н., Аде Ф.Г. Искусственный интеллект. Учебное пособие. Севастополь, 2002.

Сравнительная таблица: Список источников

Список источников	
	Учебное пособие «Основы искусственного интеллекта». Глава 9 «Обработка естественного языка (NLP)»: токенизация, век

График: иллюстрация зависимости



Иллюстративные значения (не результаты эксперимента).

Итоги и вопросы для обсуждения

1. Объясните понятие «Понятие обработки естественного языка» и приведите пример.

2. Объясните понятие «Особенности естественного языка» и приведите пример.

3. Объясните понятие «Общая схема обработки текста» и приведите пример.

4. Объясните понятие «Очистка текстовых данных» и приведите пример.

5. Объясните понятие «Токенизация» и приведите пример.

6. Объясните понятие «Нормализация текста» и приведите пример.

7. Объясните понятие «Стоп-слова» и приведите пример.