Федеральное государственное бюджетное

образовательное учреждение высшего образования

«Воронежский государственный университет»

Лабораторная работа №6

Предмет: «Введение в компьютерную лингвистику»

Доклад по теме: «Интерпретация результатов анализа.»

Работу выполнила:

Обучающаяся Подхолзина В. М.

Математический факультет

Направление: Информационно-аналитические

системы безопасности

1 курс, группа 5, подгруппа 2

Преподаватель: Донина О.В.

Воронеж 2025г.

**Оглавление**

Введение………………………………………………………………………….

Глава 1. Как дата-аналитик работает с данными……………………………...

* 1. Определение целей анализа…………………………………………...
  2. Сбор данных……………………………………………………………
  3. Предобработка данных…………………………………………………
  4. Анализ данных…………………………………………………………
  5. Интерпретация результатов……………………………………………
  6. Представление данных…………………………………………………

Вывод………………………………………………………………………

Глава 2. Основные методы представления данных……………………………

Вывод……………………………………………………………………….

Заключение…………………………………………………………………………

Список литературы…………………………………………………………………

**Введение**

В современном мире данные представляют собой один из самых ценных ресурсов. Компании и организации стремятся использовать данные для принятия обоснованных решений, повышения эффективности бизнес-процессов и создания конкурентных преимуществ. Работа дата-аналитика заключается не только в сборе и обработке данных, но и в их интерпретации, что позволяет превратить сырые данные в полезную информацию. В данном проекте мы рассмотрим основные этапы работы дата-аналитика, методы анализа и представления данных.

**Глава 1.** **Как дата-аналитик работает с данными.**

Всё начинается с понимания, что роль дата-аналитика заключается не только в работе с числами и статистикой, но и в способности рассказывать истории на основе данных. Рассмотрим подробнее, как дата-аналитик работает с данными на каждом этапе.

* 1. **Определение целей анализа**

Прежде чем приступить к работе с данными, дата-аналитик должен определить цели анализа. Это включает в себя:

- Обсуждение с заинтересованными сторонами: Дата-аналитик работает с командами (например, маркетинг, продажи, операционные), чтобы понять, какие вопросы необходимо решить.

- Постановка задач: Определение ключевых показателей эффективности (KPI) и желаемых результатов, которые помогут в оценке успеха анализа.

**1.2 Сбор данных**

**1.2.1 Источники данных**

Сбор данных может происходить из различных источников:

- Внутренние базы данных: Серверы компании, CRM-системы, ERP-системы, файлы Excel и другие формы хранения данных.

- Внешние источники: Открытые наборы данных, API других сервисов (например, Google Analytics, социальных сетей), опросы и исследования.

- Качественные данные: Обратная связь от клиентов, интервью, фокус-группы.

**1.2.2 Методы сбора данных**

- SQL-запросы: Используются для извлечения данных из реляционных баз данных.

- API: Ручной или автоматизированный сбор данных с веб-сайтов и платформ.

- Web Scraping: Использование библиотек, таких как Beautiful Soup или Scrapy, для извлечения данных с веб-страниц.

**1.3 Предобработка данных**

**1.3.1 Очистка данных**

Предобработка данных — один из наиболее критичных этапов на пути к успеху анализа:

- Удаление дубликатов: Проверка на наличие повторяющихся записей и их удаление.

- Обработка пропущенных значений: Устранение, замена средними значениями, медианами или предсказание недостающих данных.

- Исправление ошибок: Например, ненормализованные данные или опечатки.

**1.3.2 Преобразование данных**

После очистки данных необходимо их преобразовать для дальнейшего анализа:

- Нормализация и стандартизация: Приведение значений к одному масштабу.

- Агрегация: Объединение данных для получения более высокоуровневой информации (например, суммирование продаж по месяцам).

- Кодирование категориальных данных: Преобразование категориальных переменных в числовые (например, использование one-hot encoding).

**1.4 Анализ данных**

**1.4.1 Описательный анализ**

Описательная статистика помогает получить представление о данных:

- Среднее, медиана, мода: Основные показатели центральной тенденции.

- Стандартное отклонение и дисперсия: Оценка разброса данных.

- Корреляционный анализ: Выявление связей между переменными.

**1.4.2 Инференциальный анализ**

Дата-аналитик может использовать статистические тесты для выявления значимости результатов:

- t-тест: Сравнение средних значений двух групп.

- ANOVA: Сравнение средних значений более чем двух групп.

**1.4.3 Модели машинного обучения**

Если анализ требует более сложной обработки, следует использовать методы машинного обучения:

- Регрессия: Прогнозирование непрерывных значений.

- Классификация: Определение категорий на основе признаков (например, спам или не спам).

- Кластеризация: Группировка данных по схожести.

**1.5 Интерпретация результатов**

После завершения анализа результаты должны быть интерпретированы для обеспечения их понятности и полезности:

- Знания из анализа: Использование полученных данных для формулирования важных выводов.

- Связывание с бизнес-проблемами: Понимание того, как результаты влияют на бизнес-цели и задачи.

- Обсуждение с коллегами: Взаимодействие с другими отделами для проверки выводов.

**1.6 Представление данных**

**1.6.1 Визуализация данных**

Эффективная визуализация позволяет передать информацию ясно и просто:

- Графики и диаграммы: Использование линейных графиков, столбчатых диаграмм и тепловых карт.

- Интерактивные визуализации: Создание дашбордов (например, в Tableau или Power BI), где пользователи могут исследовать данные самостоятельно.

**1.6.2 Подготовка отчетов**

Дата-аналитик должен создать отчет, который включает:

- Введение с целями анализа: Краткое описание проблемы и запланированного анализа.

- Методология: Описание методов, использованных в процессе.

- Результаты и обсуждение: Основные выводы, поддержанные визуализациями.

- Рекомендации и выводы: Конкретные шаги, которые компания может предпринять на основе анализа.

**Вывод:**

Работа дата-аналитика с данными представляет собой сложный и многоэтапный процесс, включающий сбор, обработку, анализ, интерпретацию и представление данных. Каждая из этих стадий требует различных навыков и знаний, а успех анализа зависит от умения аналитика находить нужные данные, извлекать из них полезную информацию и доносить её до заинтересованных сторон.

**Глава 2. Основные методы представления данных.**

Методы представления данных играют ключевую роль в том, как информация воспринимается и интерпретируется пользователями. Эффективная визуализация данных помогает быстро понять основные тенденции, выявить закономерности и проанализировать результаты исследований. Ниже представлены основные методы представления данных, их описание и примеры использования.

**1. Таблицы**

Таблицы являются одним из самых простых и удобных способов представления данных. Они позволяют компактно организовать информацию и обеспечить легкий доступ к конкретным значениям.

Преимущества:

- Компактность: позволяют видеть большие объемы данных одновременно.

- Четкость: данные организованы по строкам и колонкам, что упрощает восприятие.

Когда использовать:

- Когда необходимо представить многочисленные данные с конкретными значениями (например, финансовые отчеты, результаты опросов).

**2. Гистограммы**

Гистограммы используются для отображения распределения числовых данных. Они показывают частоту значений в заданных интервалах (или бингах).

Преимущества:

- Легкость восприятия распределения данных.

- Видимость закономерностей, таких как асимметрия и наличие выбросов.

Когда использовать:

- При необходимости показать распределение непрерывных данных (например, рост, доход, возраст).

Пример: Гистограмма, показывающая распределение возраста клиентов.

**3. Столбчатые диаграммы**

Столбчатые диаграммы (или бар-чарты) предназначены для сравнения категориальных данных. Каждая категория представляется вертикальной (или горизонтальной) полосой, длина которой соответствует величине значения.

Преимущества:

- Простота и наглядность.

- Хорошо подходит для сравнения данных между разными группами.

Когда использовать:

- Когда необходимо сравнить несколько категорий (например, продажи по регионам).

Пример: Столбчатая диаграмма, показывающая продажи различных продуктов.

**4. Линейные графики**

Линейные графики показывают изменения значений во времени, что делает их полезными для анализа трендов и динамики.

Преимущества:

- Позволяет легко увидеть тренд и предсказать дальнейшее поведение.

- Можно использовать для сопоставления нескольких временных рядов.

Когда использовать:

- Для визуализации данных, собранных по временной шкале (например, рост дохода по месяцам).

Пример: Линейный график, показывающий изменения продаж за последние 12 месяцев.

**5. Круговые диаграммы**

Круговые диаграммы (или пай-чарты) отображают пропорциональное распределение категорий от общего количества. Каждая часть круга представляет собой долю от всего объема данных.

Преимущества:

- Хорошо передает идею долей и пропорций.

- Визуально привлекательно.

Когда использовать:

- Когда необходимо показать, как различные части составляют целое (например, доля рынка различных компаний).

Пример: Круговая диаграмма, показывающая распределение расходов по категориям.

**6. Тепловые карты**

Тепловые карты используют цвет для представления данных в двумерном виде, что позволяет быстро выявлять закономерности и аномалии.

Преимущества:

- Компактность информации и наглядность.

- Легкость в выявлении закономерностей и корреляций.

Когда использовать:

- При необходимости визуализировать большие объемы данных, например, для сравнительного анализа различных показателей.

Пример: Тепловая карта, показывающая производительность различных регионов по месяцам.

**7. Ассоциативные карты**

Этот метод позволяет визуализировать связи между различными данными в виде узлов и линий, где каждый узел представляет собой концепцию, а линии показывают отношения между ними.

Преимущества:

- Визуальная ясность взаимосвязей между элементами.

- Позволяет легко показать иерархию данных.

Когда использовать:

- Для представления денежных потоков, связей между переменными и концепциями.

Пример: Ассоциативная карта, показывающая связи между различными показателями бизнеса.

**8. Дашборды**

Дашборды представляют собой сводные визуализации, предоставляющие полный обзор ключевых показателей (KPI) в одном месте. Они могут включать различные типы визуализаций, такие как графики, таблицы и карты.

Преимущества:

- Интуитивность: способствуют быстрому взгляду на текущую ситуацию.

- Интерактивные элементы: позволяют пользователям углубляться в данные.

Когда использовать:

- Для мониторинга KPI бизнеса в реальном времени (например, дашборды для финансового анализа).

Пример: Дашборд, показывающий данные о продажах, затратах и прибыли компании в режиме реального времени.

**9. Географические карты**

Географические карты отображают данные, связанные с географическими регионами. Они могут использоваться для визуализации не только местоположений, но и значений различных показателей, связанных с этими местоположениями.

Преимущества:

- Позволяют видеть пространственные корреляции и распределение данных.

- Ясно показывают, какие регионы имеют лучшие или худшие показатели.

Когда использовать:

- Для анализа распределения данных по регионам (например, продаж в разных странах).

Пример: Карта, показывающая уровень продаж по штатам или регионам.

**10. Инфографика**

Инфографика — комбинация текста и визуальных элементов, позволяющая кратко и наглядно представить данные и их контекст.

Преимущества:

- Привлекательность: способствует удержанию внимания.

- Всестороннее представление сложных данных.

Когда использовать:

- Для представления больших объемов данных в понятном и интересном формате (например, для учебных материалов или отчетов для СМИ).

Пример: Инфографика, демонстрирующая результаты опроса клиентов с использованием диаграмм, иконок и текстов.

**Вывод:**

Выбор метода представления данных зависит от целей анализа, типа данных и целевой аудитории. Эффективные визуализации помогают не только в передаче информации, но и в принятии обоснованных решений. Использование разнообразных методов позволяет адаптировать представление данных к различным контекстам и задачам, обеспечивая более глубокое понимание и интерпретацию полученных результатов.

**Заключение**

Работа дата-аналитика включает множество этапов — от сбора и обработки данных до анализа и интерпретации результатов. Умение представлять данные в визуально понятной форме является необходимым навыком для распространения инсайтов и рекомендаций среди заинтересованных сторон. Дата-аналитика является неотъемлемой частью современного бизнеса, и его роль продолжает расти в связи с увеличением объемов данных и потребностью в их интерпретации.

**Список литературы**

1. https://practicum.yandex.ru/blog/professiya-analitik-dannyh/

2. https://ahaslides.com/ru/blog/10-methods-of-data-presentation/

3. <https://practicum.yandex.ru/data-analyst/?var=daan105&utm_source=yandex&utm_medium=cpc&utm_campaign=Yan_Sch_RF_Data_dataAn_b2c_Gener_Regular_1_460&utm_content=sty_search%3As_none%3Acid_109820554%3Agid_5436137655%3Apid_51253443914%3Aaid_16072808703%3Acrid_0%3Arid_51253443914%3Ap_1%3Apty_premium%3Amty_syn%3Amkw_%3Adty_desktop%3Acgcid_26898098%3Arn_Воронеж%3Arid_193&utm_term=аналитика+данных+методы&yclid=12779454828821610495>

Список вопросов:

1. Какова основная роль дата-аналитика?

2. Почему важно определение целей анализа перед началом работы с данными?

3. Какие факторы влияют на выбор источников данных для анализа?

4. Каковы основные этапы работы дата-аналитика с данными?

5. Почему методы представления данных играют ключевую роль в восприятии информации?

6. Как визуализация данных может повлиять на процесс принятия решений?

7. Какие проблемы могут возникнуть, если данные неверно представлены?

8. Что такое гистограмма и как она работает?

9. Каковы основные преимущества использования гистограмм для отображения данных?

10. Почему линейные графики являются хорошим выбором для отображения временных рядов?

**Кроссворд**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  | **7** |  |  |  |  | **6** |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | **3** |  |  |  |  |  |  |  |  |  |
| **4** |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  | **2** |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  | **1** |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  | **5** |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |

1. **Процесс систематического анализа данных с целью выявления закономерностей, трендов и определённых фактов**, которые могут помочь в принятии решений.

2. **Графическое представление данных линейными отрезками или геометрическими фигурами**, позволяющее быстро оценить соотношение нескольких величин.

3. Действие, направленное на изменение свойств предмета труда при выполнении технологического процесса.

4. Он предполагает определённую последовательность действий на основе чётко осознанного плана.

5. **Метод исследования, характеризующийся выделением и изучением отдельных частей объектов исследования.**

6. Объект реальной действительности, который даёт возможность получить информацию о чём-либо.

7. сообщения (числа, текст, изображения, звук, видео и др.), представленные в виде, допускающем их обработку программно-аппаратными средствами или интерпретацию человеком.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  | **7** |  |  |  |  | **6** |  |  |
|  |  |  |  |  |  | д |  |  |  |  | и |  |  |
|  |  |  |  |  |  | а |  |  |  |  | с |  |  |
|  |  |  |  |  |  | н |  |  |  |  | т |  |  |
|  |  |  |  | **3** |  | н |  |  |  |  | о |  |  |
| **4** | м | е | т | о | д | ы |  |  |  |  | ч |  |  |
|  |  |  |  | б |  | е |  | **2** |  |  | н |  |  |
|  |  |  |  | р |  |  |  | д |  |  | и |  |  |
|  |  |  | **1** | а | н | а | л | и | т | и | к | а |  |
|  |  |  |  | б |  |  |  | а |  |  |  |  |  |
|  |  |  |  | о |  |  |  | г |  |  |  |  |  |
|  |  |  |  | т |  |  |  | р |  |  |  |  |  |
|  |  |  |  | к |  |  | **5** | а | н | а | л | и | з |
|  |  |  |  | а |  |  |  | м |  |  |  |  |  |
|  |  |  |  |  |  |  |  | м |  |  |  |  |  |
|  |  |  |  |  |  |  |  | а |  |  |  |  |  |

Лабораторная работа по теме «Интерпретация результатов анализа.»

**Введение**

Компьютерная лингвистика — это междисциплинарная область, объединяющая лингвистику и информатику, направленная на автоматическую обработку и анализ естественного языка. Одним из ключевых этапов в этой области является интерпретация результатов автоматического анализа текста, которая позволяет понять смысл, структуру и особенности языковых данных.

Цель данной лабораторной работы — научиться правильно интерпретировать результаты автоматического анализа текста, полученные с помощью различных инструментов и методов, а также выявлять возможные ошибки и области для улучшения.

**Основные этапы анализа текста**

Перед интерпретацией результатов важно понимать, какие виды анализа выполняются:

* Морфологический анализ — определение частей речи, морфологических характеристик слов.
* Синтаксический разбор — построение синтаксической структуры предложения.
* Семантический анализ — выявление смысловых связей и значений.
* Анализ связности и дискурса — понимание контекста и связности текста.

**Интерпретация результатов**

**1. Морфологический анализ**

Что важно:

* Проверить правильность определения частей речи.
* Обратить внимание на морфологические признаки (число, род, падеж и т.д.).
* Выявить ошибки, например, неправильное определение части речи или морфологических характеристик.

Пример:

Если слово "бегает" определено как существительное, это ошибка, так как правильная часть речи — глагол.

**2. Синтаксический разбор**

Что важно:

* Анализировать структуру предложения.
* Проверить правильность построения синтаксических связей.
* Обратить внимание на ошибки в связях, такие как неправильное определение подлежащего или сказуемого.

Пример:

В предложении "Мальчик читает книгу" правильная структура — "Мальчик" (подлежащее), "читает" (сказуемое), "книгу" (дополнение). Если разбор показывает иначе, необходимо проверить корректность анализа.

**3. Семантический анализ**

Что важно:

* Оценить правильность определения