

**Федеральное государственное автономное образовательное  
учреждение высшего образования  
«Московский физико-технический институт  
(национальный исследовательский университет)»**

**УТВЕРЖДЕНО**

**Директор физтех-школы  
прикладной математики и  
информатики**

**А.М. Райгородский**

	<b>Рабочая программа дисциплины (модуля)</b>
<b>по дисциплине:</b>	Сбор и разметка данных
<b>по направлению:</b>	Прикладная математика и информатика
<b>профиль подготовки:</b>	А1360: Передовые методы искусственного интеллекта Физтех-школа Прикладной Математики и Информатики центр практик и стажировок ФПМИ
<b>курс:</b>	3
<b>квалификация:</b>	бакалавр

Семестр, формы промежуточной аттестации: 5 (осенний) - Дифференцированный зачет

Аудиторных часов: 60 всего, в том числе:

лекции: 30 час.

семинары: 30 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 30 час.

Всего часов: 90, всего зач. ед.: 2

Программу составил: А.Ю. Ширяев, заведующий лабораторией

Программа обсуждена на заседании центра практик и стажировок ФПМИ 12.02.2024

## Аннотация

Дисциплина "Сбор и разметка данных" представляет собой введение в методы, инструменты и технологии, используемые для сбора, обработки и разметки данных в контексте анализа и машинного обучения. Студенты изучат основные принципы сбора данных из различных источников, таких как базы данных, веб-сайты, датчики и другие устройства. Кроме того, они узнают о методах разметки данных, включая разметку текстов, изображений и аудио- и видеофайлов. Дисциплина также охватывает вопросы качества данных, этические аспекты сбора и использования информации, а также автоматизацию процессов сбора и разметки данных. В результате изучения студенты смогут применять полученные знания для создания надежных наборов данных, необходимых для обучения моделей машинного обучения и анализа данных в различных областях.

### 1. Цели и задачи

#### Цель дисциплины

Обучить студентов основам сбора и разметки данных для машинного обучения и искусственного интеллекта. Развить у студентов практические навыки работы с различными источниками данных, методами их сбора и подготовки к обучению моделей. Способствовать развитию у студентов критического мышления и способности оценивать качество данных, их релевантность и влияние на эффективность обучения моделей.

#### Задачи дисциплины

- развитие практических навыков сбора и разметки данных;
- развитие аналитических навыков;
- развить способность оценивать качество данных, их релевантность и влияние на эффективность обучения модели.

### 2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-1 Способен применять фундаментальные знания, полученные в области физико-математических и (или) естественных наук и использовать их в профессиональной деятельности	ОПК-1.1 Способен анализировать поставленную задачу, намечать пути ее решения
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области

### 3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основные понятия сбора и разметки данных;
- различные источники данных и методы их сбора;
- инструменты и техники разметки данных;
- этические аспекты сбора и разметки данных.

уметь:

- собрать данные из различных источников;
- разметить данные для обучения модел;
- подготовить данные к обучению модели;
- оценивать качество данных и их влияние на эффективность обучения.

владеть:

- практическими навыками сбора и разметки данных;
- способностью решать проблемы, связанные с данными;
- пониманием этических аспектов сбора и разметки данных.

#### 4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

##### 4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Основы сбора данных	5	5		5
2	Этика и законность сбора данных	5	5		5
3	Инструменты и технологии для сбора данных	5	5		5
4	Методы разметки данных	5	5		5
5	Качество данных и оценка достоверности	10	10		10
Итого часов		30	30		30
Подготовка к экзамену		0 час.			
Общая трудоёмкость		90 час., 2 зач.ед.			

##### 4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 5 (Осенний)

###### 1. Основы сбора данных

Источники данных: обзор основных источников данных, таких как опросы, интервью, наблюдения, базы данных и др. Методы сбора данных: сравнение различных методов сбора данных, включая качественные и количественные подходы.

###### 2. Этика и законность сбора данных

Конфиденциальность и защита данных: важность обеспечения конфиденциальности данных и соответствия законодательству о защите персональных данных. Этические аспекты сбора данных: обсуждение этических проблем, связанных с сбором и использованием данных.

###### 3. Инструменты и технологии для сбора данных

Автоматизированный сбор данных: обзор инструментов для автоматизации процесса сбора данных, таких как веб-скрейпинг, API и др. Инструменты для опросов и интервью: рассмотрение программного обеспечения для проведения опросов и интервью.

###### 4. Методы разметки данных

Ручная разметка: техники и инструменты для ручной разметки данных. Машинное обучение и разметка данных: применение методов машинного обучения для автоматизации процесса разметки данных.

###### 5. Качество данных и оценка достоверности

Оценка качества данных: методы оценки достоверности и качества собранных данных.  
Управление ошибками при сборе и разметке данных: стратегии управления ошибками и неточностями в данных.

## **5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)**

Стандартная учебная аудитория, оборудованная компьютером и мультимедийным оборудованием (проектор, звуковая система) для докладов и презентаций.

## **6. Перечень рекомендуемой литературы**

Основная литература

1. Python и анализ данных, Электрон. версия печ. публикации / У. Маккини. — Москва, ДМК Пресс, 2020

Дополнительная литература

## **7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)**

Не используются

## **8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)**

На занятиях используется компьютер и мультимедийное оборудование (проектор, звуковая система),

## **9. Методические указания для обучающихся по освоению дисциплины (модуля)**

Студент, изучающий дисциплину, должен с одной стороны, овладеть общим понятийным аппаратом, а с другой стороны, должен научиться применять теоретические знания на практике.

Успешное освоение дисциплины требует:

- посещения студентом всех видов аудиторных занятий;
- качественной самостоятельной подготовки к практическим занятиям, активной работы на них;
- активной самостоятельной и аудиторной работы студента;
- своевременной сдачи преподавателю заданий по аудиторным видам работ.

**ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)**

<b>по направлению:</b>	Прикладная математика и информатика
<b>профиль подготовки:</b>	АІ360: Передовые методы искусственного интеллекта Физтех-школа Прикладной Математики и Информатики центр практик и стажировок ФПМИ
<b>курс:</b>	<u>3</u>
<b>квалификация:</b>	бакалавр
Семестр, формы промежуточной аттестации: 5 (осенний) - Дифференцированный зачет	
<b>Разработчик:</b>	А.Ю. Ширяев, заведующий лабораторией

## 1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-1 Способен применять фундаментальные знания, полученные в области физико-математических и (или) естественных наук и использовать их в профессиональной деятельности	ОПК-1.1 Способен анализировать поставленную задачу, намечать пути ее решения
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области

## 2. Показатели оценивания компетенций

В результате изучения дисциплины «Сбор и разметка данных» обучающийся должен:

### знать:

- основные понятия сбора и разметки данных;
- различные источники данных и методы их сбора;
- инструменты и техники разметки данных;
- этические аспекты сбора и разметки данных.

### уметь:

- собрать данные из различных источников;
- разметить данные для обучения модел;
- подготовить данные к обучению модели;
- оценивать качество данных и их влияние на эффективность обучения.

### владеть:

- практическими навыками сбора и разметки данных;
- способностью решать проблемы, связанные с данными;
- пониманием этических аспектов сбора и разметки данных.

## 3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

1. Какие основные методы сбора данных вы знаете? Приведите примеры ситуаций, когда каждый из них может быть наиболее эффективен.
2. Какие этические проблемы могут возникнуть при сборе и разметке данных? Какие меры могут быть предприняты для защиты конфиденциальности и обеспечения этичного использования данных?
3. Каким образом автоматизированный сбор данных отличается от ручного? Какие преимущества и недостатки у каждого из этих методов?
4. Какие инструменты вы бы использовали для проведения опросов и интервью? Обсудите их особенности и возможности.
5. Что такое разметка данных? Какие методы разметки данных вы знаете, и какие сценарии использования подходят для каждого из них?
6. Какие технологии машинного обучения могут быть использованы для автоматизации процесса разметки данных? Приведите примеры задач, в которых это может быть особенно полезно.
7. Как можно оценить качество собранных данных? Какие методы оценки достоверности данных вы бы применили в конкретной ситуации?
8. Какие ошибки и неточности могут возникнуть при сборе и разметке данных, и как их можно управлять? Приведите примеры стратегий управления ошибками в данных.
9. Какие законы и стандарты регулируют сбор и использование персональных данных? Какие требования они предъявляют к процессам сбора и разметки данных?

10. Каким образом сбор и разметка данных влияют на качество и точность анализа данных? Предоставьте примеры из практических областей, где правильная сборка и разметка данных играют ключевую роль.

#### **4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся**

1. Какие методы сбора и разметки данных наиболее эффективны при работе с неструктурированными данными, такими как текст или изображения?
2. Какие трудности могут возникнуть при сборе данных из различных источников? Какие стратегии можно использовать для обеспечения единообразия и качества данных?
3. Какие факторы следует учитывать при выборе обучающей выборки для модели машинного обучения? Как можно оценить, достаточно ли данных для обучения модели?
4. Какие методы можно применить для обнаружения и устранения выбросов и аномалий в данных в процессе сбора и разметки?
5. Каким образом сбор и разметка данных могут быть автоматизированы с использованием технологий и инструментов, таких как RPA (Robotic Process Automation) или инструменты для разметки данных?
6. Какие могут быть последствия недостаточной разметки данных для обучения моделей машинного обучения? Как избежать проблем, связанных с неправильной или неполной разметкой?
7. Какие методы можно применить для защиты данных от утечек и несанкционированного доступа в процессе их сбора и разметки?
8. Каким образом сбор и разметка данных могут быть оптимизированы для повышения производительности и качества процесса обучения моделей машинного обучения?
9. Какие роли могут играть люди и алгоритмы в процессе сбора и разметки данных? Как можно эффективно сочетать человеческий труд и автоматизацию в этом процессе?
10. Каким образом сбор и разметка данных влияют на конечные результаты анализа данных и принятие решений в бизнесе? Какие практические примеры успешного использования данных можно привести в этом контексте?

#### **Критерии оценивания**

Оценка "Отлично" (10) - полностью и вовремя решены все задачи без ошибок. Продemonстрирован грамотный подход к решению задач, реализованы оптимальные алгоритмы, код оформлен в едином удобочитаемом стиле.

Оценка "Отлично" (9) - полностью и вовремя решены все задачи без ошибок. Продemonстрирован грамотный подход к решению задач, реализованы оптимальные алгоритмы.

Оценка "Отлично" (8) - полностью и вовремя решены все задачи без ошибок. Продemonстрирован грамотный подход к решению задач.

Оценка "Хорошо" (7) - полностью решены все задачи. Допущены несущественные ошибки.

Оценка "Хорошо" (6) - полностью решено большинство задач. В некоторых задачах допущены и не исправлены ошибки, либо некоторые задачи решены частично.

Оценка "Хорошо" (5) - полностью решено две трети задач. В некоторых задачах допущены и не исправлены ошибки, либо некоторые задачи решены частично.

Оценка "Удовлетворительно" (4) - полностью решено более половины задач. В остальных задачах допущены и не исправлены ошибки, либо некоторые задачи решены частично.

Оценка "Удовлетворительно" (3) - полностью решено более половины задач.

Оценка "Неудовлетворительно" (2) - решено менее половины задач.

Оценка "Неудовлетворительно" (1) - не решено ни одной задачи.

#### **5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности**

Дифференцированный зачет может проводиться по итогам текущей успеваемости и сдачи заданий и других видов работ, предусмотренных программой дисциплины и (или) путем организации специального опроса, проводимого в устной и (или) письменной форме.

При проведении устного дифференцированного зачета обучающемуся предоставляется 30 минут на подготовку. Опрос обучающегося не должен превышать одного астрономического часа.

Во время проведения дифференцированного зачета обучающиеся могут пользоваться программой дисциплины, а также справочной литературой, конспектами лекций или другими материалами.