
УТВЕРЖДЕНА

Решением Ученого совета
АНО ВО «Центральный университет»
«25» мая 2026 г.
Протокол № 2

**Рабочая программа дисциплины (модуля)
«Разметка наборов данных. Авито»**

Направление подготовки: 02.04.01 Математика и компьютерные науки

Направленность (профиль) подготовки: Продуктовая аналитика

Квалификация (степень) выпускника: магистр

Форма обучения: очная

Срок освоения программы: 2 года

Год набора: 2026

**Москва
2026**

Содержание

1. Краткая характеристика дисциплины (модуля)	3
2. Перечень планируемых результатов обучения.....	3
3. Тематический план.....	7
4. Содержание дисциплины (модуля).....	7
5. Учебно-методическое обеспечение	8
6. Материально-техническое обеспечение	8
7. Методические и оценочные материалы	10

1. Краткая характеристика дисциплины (модуля)

Рабочая программа дисциплины (модуля) «Разметка наборов данных. Авито» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – магистратура по направлению 02.04.01 Математика и компьютерные науки, профиль Продуктовая аналитика, утвержденный приказом Министерства науки и высшего образования Российской Федерации № 810 от 23.08.2017 года.

Изучение дисциплины (модуля) «Разметка наборов данных. Авито» позволяет сформировать у студентов критически важный навык подготовки качественных данных для машинного обучения.

Место дисциплины (модуля) в структуре образовательной программы

Настоящая дисциплина (модуль) включена в учебный план по программе подготовки магистратуры по направлению 02.04.01 Математика и компьютерные науки, профиль Продуктовая аналитика и входит в вариативную часть Блока 1, формируемую участниками образовательных отношений как дисциплина по выбору.

Дисциплина (модуль) изучается на 1 или 2 курсе во 2, 3 или 4 семестре на выбор, доступна для прохождения при условии успешного завершения дисциплины (модуля) «Основы Python».

Цель изучения дисциплины (модуля): освоение технологий эффективной подготовки датасетов для задач машинного обучения, включая выбор стратегий разметки, работу с краудсорсинговыми платформами и применение больших языковых моделей для автоматизации процесса.

Задачи изучения дисциплины (модуля):

- раскрыть принципы организации краудсорсинга и настройки проектов на платформах разметки;
- сформировать навык выбора между ручной, автоматической и гибридной разметкой в зависимости от задачи;
- освоить методы оптимизации разметки через активное обучение и полуавтоматические подходы;
- изучить технологии генерации данных: аугментации, 3D-моделирование и генеративные модели;
- отработать практику применения больших языковых моделей для разметки текстов и создания инструкций;
- развить умение оценивать качество датасетов и соотносить затраты на разметку с бизнес-ценностью модели.

В результате освоения дисциплины (модуля) обучающийся должен:

знать:

- основные понятия и виды краудсорсинга, включая их применение в разметке данных для задач машинного обучения;
- принципы и методы ручной, автоматической и гибридной разметки данных;
- сущность и практическое применение эвристик, активного обучения и полуавтоматического обучения (semi-supervised learning) в оптимизации разметки;
- источники открытых данных, методы их хранения и подготовки для разметки;
- основы генерации данных: аугментации, 3D-моделирование, а также принципы работы генеративных моделей (GAN, диффузионные модели);
- фундаментальные принципы работы больших языковых моделей: базовое обучение (pretrain), контролируемая донастройка (SFT), RLHF (обучение с подкреплением по обратной связи от людей), рассуждение (reasoning), методы дообучения;

- сравнительные характеристики разметки с помощью краудсорсинга и LLM, включая особенности применения в условиях Авито.

уметь:

- создавать и настраивать проекты на краудсорсинговых платформах (на примере Яндекс.Задания и Толоки): писать инструкции, разрабатывать обучение и контроль, загружать тестовые пулы и анализировать результаты разметки;
- применять эвристики для автоматической разметки на основе свойств объектов и внешних данных;
- использовать активное обучение и полуавтоматическое обучение (semi-supervised для снижения трудозатрат при ручной разметке;
- генерировать дополнительные обучающие примеры для задач компьютерного зрения и NLP с помощью аугментаций, 3D-моделирования и генеративных моделей;
- формулировать промпты для разметки с помощью LLM и сравнивать качество таких данных с краудсорсинговыми;
- оценивать и выбирать оптимальные инструменты и подходы к разметке в зависимости от задачи, бюджета и требований к качеству;
- подготавливать чистые, разнообразные и репрезентативные дата-сеты для обучения моделей.

владеть:

- платформами краудсорсинга: Яндекс.Задания, Толока;
- инструментами для автоматической разметки и предразметки (включая эвристические скрипты и правила);
- методами активного обучения и полуавтоматического обучения (semi-supervised learning); для оптимизации процесса;
- генеративными технологиями: аугментации изображений, GAN, диффузионные модели;
- LLM для генерации и разметки текстовых данных с использованием промптов.
- подходами к ценообразованию и сравнению эффективности различных инструментов разметки;
- практическими шаблонами создания дата-сетов: от сбора данных до финальной валидации и балансировки.

2. Перечень планируемых результатов обучения

Компетенции, формируемые в результате освоения дисциплины (модуля) при проведении учебных занятий в форме контактной работы обучающихся с педагогическими работниками Университета и в форме самостоятельной работы обучающихся:

Компетенция	Содержание компетенции	Индикатор компетенции	Перечень планируемых результатов обучения по дисциплине (модулю)
УК-6.	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки	УК-6.1.	Знает основные методы самооценки и анализа своей деятельности, а также принципы управления временем и целеполагания
		УК-6.2.	Умеет ставить реалистичные и достижимые цели, определять приоритеты в своей деятельности, а также разрабатывать и внедрять планы по совершенствованию своих навыков и компетенций на основе полученной самооценки
		УК-6.3.	Имеет практический опыт применения методов самооценки в своей профессиональной деятельности, включая участие в тренингах, семинарах и проектах, направленных на развитие личной эффективности и профессионального роста
ОПК-1.	Способен находить, формулировать и решать актуальные и значимые проблемы прикладной и компьютерной математики	ОПК-1.1.	Знает основные методы и подходы к решению задач прикладной и компьютерной математики, включая алгоритмы, математическое моделирование и теорию оптимизации, а также современные инструменты и технологии, используемые в этой области
		ОПК-1.2.	Умеет анализировать и формулировать математические задачи, применять соответствующие методы и алгоритмы для их решения, а также интерпретировать и представлять результаты в

			понятной и доступной форме
		ОПК-1.3.	Имеет практический опыт работы над проектами или исследованиями в области прикладной и компьютерной математики, включая участие в конкурсах, олимпиадах или научных публикациях, где были решены актуальные и значимые задачи
ПК-3.	Способен решать задачи профессиональной деятельности в области продуктовой аналитики, формулировать результаты анализа и выявлять последствия полученных данных для принятия обоснованных решений и оптимизации продуктов	ПК-3.1.	Знает методы и инструменты продуктовой аналитики
		ПК-3.2.	Умеет применять аналитические инструменты и программное обеспечение для обработки и визуализации данных, а также формулировать выводы на основе проведенного анализа
		ПК-3.3.	Имеет опыт работы над реальными проектами в области продуктовой аналитики, включая анализ пользовательского поведения и оптимизацию продуктов на основе полученных данных

3. Тематический план

№ п/п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы				ТКУ (текущий контроль успеваемости)
		Очная форма				
		Аудиторная работа		Контроль	Самостояте льная работа	
		Лекции	Семинары (практичес кие занятия)			
1	Краудсорсинг	10	10		42	Домашние задания
2	Упрощение разметки	10	10		42	Домашние задания
3	Генерация данных	10	10		42	Домашние задания
	Экзамен			4		
Итого:		30	30	4	126	
Объем дисциплины (модуля) (в ак. ч.)		190				
Объем дисциплины (модуля) (в зач. ед.)		5				

4. Содержание дисциплины (модуля)

№п/п	Наименование раздела дисциплины (модуля)	Содержание дисциплины (модуля) по темам
1	Краудсорсинг	Основы краудсорсинга и его виды Создание датасетов с помощью краудсорсинга Основные принципы и методы разметки данных Работа с ресурсами краудсорсинга
2	Упрощение разметки	Автоматическая разметка с использованием свойств объекта Разметка с использованием сторонних данных Разметка с привлечением пользователей продукта Оптимизация ручной разметки - активное обучение Оптимизация ручной разметки - полуавтоматическое обучение (semi-supervised learning)
3	Генерация данных	Работа с данными Разметка Генерация GAN, диффузия и другие генеративные модели Введение в LLM Разметка с помощью промтов

5. Учебно-методическое обеспечение

Университет располагает полным набором лицензионного и свободно распространяемого программного обеспечения, включая продукты отечественного производства.

Каждый студент в течение всего периода обучения получает индивидуальный неограниченный доступ к электронно-библиотечной системе и электронной информационно-образовательной среде университета. Эти системы предоставляют возможность доступа к ресурсам из любой точки, где есть подключение к сети Интернет, как на территории университета, так и за его пределами.

Студентам обеспечен удаленный доступ к современным профессиональным базам данных и информационным справочным системам.

Основная литература:

1. Малов, Д. А. Глубокое обучение и анализ данных. Практическое руководство : практическое руководство / Д. А. Малов. - Санкт-Петербург : БХВ-Петербург, 2023. - 272 с. - ISBN 978-5-9775-1172-8. - Текст : электронный. - URL: <https://znanium.com/catalog/product/2123365>.

2. Лесковец, Ю. Анализ больших наборов данных : практическое руководство / Д. Дж. Ульман, Ю. Лесковец, А. Раджараман ; пер. с англ. А. А. Слинкина. - 2-е изд. - Москва : ДМК Пресс, 2023. - 500 с. - ISBN 978-5-89818-304-2. - Текст : электронный. - URL: <https://znanium.ru/catalog/product/2102592>.

Дополнительная литература:

1. Маунт, Д. Погружение в аналитику данных : практическое руководство / Д. Маунт. - Санкт-Петербург : БХВ-Петербург, 2023. - 224 с. - ISBN 978-5-9775-6866-1. - Текст : электронный. - URL: <https://znanium.com/catalog/product/2123377>.

6. Материально-техническое обеспечение

Университет располагает материально-технической базой, соответствующей действующим противопожарным правилам и нормам и обеспечивающей проведение всех видов дисциплинарной и междисциплинарной подготовки, практической и научно-исследовательской работ обучающихся, предусмотренных учебным планом.

Помещения, которые представляют собой учебные аудитории для проведения занятий лекционного типа, занятий семинарского (практического) типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы и помещения для хранения и профилактического обслуживания учебного оборудования. Помещения укомплектованы специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Изучение дисциплины (модуля) обеспечивается в учебных аудиториях, оснащенных:

- столами и стульями;
- компьютерной техникой;
- механическими калькуляторами;
- специализированным оборудованием, включая демонстрационное оборудование.

Помещения для самостоятельной работы обучающихся, в том числе приспособленные для использования инвалидами и лицами с ограниченными возможностями здоровья, оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду Университета.

Обучающимся предоставляется доступ (в том числе удаленный) к ресурсам информационно-телекоммуникационной сети «Интернет», электронным ресурсам (в том числе электронным библиотечным системам, современным профессиональным базам данных и информационным справочным системам):

№	Наименование портала (издания, курса, документа)	Ссылка
1.	Научная электронная библиотека elibrary.ru библиотека	https://elibrary.ru/defaultx.asp
2.	База данных для IT-специалистов	https://habr.com
3.	База данных ScienceDirect	https://www.sciencedirect.com
4.	Официальный сайт Министерства науки и высшего образования Российской Федерации	https://minobrnauki.gov.ru/
5.	Федеральный портал «Российское образование»	https://www.edu.ru/
6.	Информационная система "Единое окно доступа к образовательным ресурсам"	http://window.edu.ru/
7.	Единая коллекция цифровых образовательных ресурсов	http://school-collection.edu.ru/
8.	Федеральный центр информационно - образовательных ресурсов	http://fcior.edu.ru/

Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), в том числе комплект лицензионного программного обеспечения, современные профессиональные базы данных и информационные справочные системы:

Наименование ПО	Производство	Лицензионное / свободно распространяемое
Операционные системы:		
Microsoft Imagine (Windows Client, Server)	зарубежное	лицензионное
Браузеры:		
Яндекс.Браузер	отечественное	свободно распространяемое
Google Chrome	зарубежное	свободно распространяемое
Офисные приложения:		
Microsoft Imagine (Visio, OneNote)	зарубежное	лицензионное
TeXstudio	зарубежное	свободно распространяемое
Adobe Acrobat Reader	зарубежное	свободно распространяемое
Программное обеспечение для планирования и учета времени:		
Toggle app	зарубежное	свободно распространяемое
Системы управления проектами:		
Microsoft Imagine (Project)	зарубежное	лицензионное
Системы управления базами данных:		
Microsoft Imagine (SQL Server)	зарубежное	лицензионное
Системы резервного копирования (backup):		
Acronis Backup Advanced for HyperV	зарубежное	лицензионное
Справочно-правовые системы:		
КонсультантПлюс: справочно-правовая система	отечественное	лицензионное
Средства антивирусной защиты:		
Kaspersky Endpoint Security для бизнеса Стандартный Russian Edition	отечественное	лицензионное

Среды разработки:		
Visual Studio Code	зарубежное	свободно распространяемое
Bash (Unix shell)	зарубежное	свободно распространяемое
Anaconda	зарубежное	свободно распространяемое
Robotic Operating System	зарубежное	свободно распространяемое
CopelliaSim	зарубежное	свободно распространяемое
Google Colaboratory	зарубежное	свободно распространяемое
Пакеты программных средств и библиотек:		
AutoPsy	зарубежное	свободно распространяемое
Interactive Disassembler (IDA)	зарубежное	свободно распространяемое
Системы управления библиографической информацией:		
Zotero	зарубежное	свободно распространяемое
Сервисы и службы:		
Bind	зарубежное	свободно распространяемое
Docker	зарубежное	свободно распространяемое

7. Методические и оценочные материалы

Методические указания для обучающихся по освоению дисциплины (модуля)

В процессе изучения дисциплины (модуля) «Разметка наборов данных. Авито» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, семинары, домашние задания, а также различные виды самостоятельной работы обучающихся по заданию преподавателя, направленные на развитие навыков профессиональной лексики, закрепление практических профессиональных компетенций, поощрение инициатив.

Лекция – систематическое, последовательное, монологическое изложение преподавателем учебного материала, как правило, теоретического характера.

В процессе лекций рекомендуется вести конспект лекций: кратко и схематично фиксировать основные идеи, выводы и обобщения лекции; выделять важные мысли, ключевые слова и термины. Необходимо отметить вопросы или материалы, которые вызывают затруднения, и попытаться найти ответы в рекомендованной литературе. Если разобраться в материале не удастся, следует сформулировать вопрос и задать его преподавателю на консультации или во время семинарского (практического) занятия.

Семинар — это форма учебной деятельности, проводимая в учебном заведении под руководством преподавателя, где студенты активно участвуют в обсуждениях, практических заданиях и других формах взаимодействия.

Для успешной подготовки к семинару рекомендуется заранее ознакомиться с темой занятия и основными материалами, чтобы иметь возможность активно участвовать в обсуждении. Также полезно подготовить вопросы и идеи для обсуждения, что поможет глубже понять материал и продемонстрировать заинтересованность.

Домашнее задание – набор задач по темам недели.

При работе над домашними заданиями важно внимательно ознакомиться с требованиями и сроками выполнения. Рекомендуется разбивать задания на этапы, чтобы избежать перегрузки и лучше усвоить материал, использовать различные источники информации, включая учебники и онлайн-ресурсы, для более глубокого понимания темы.

Бонусные баллы — это оценки, которые студенты могут получить за выполнение дополнительных заданий.

Формат бонусных баллов позволяет студентам улучшить общую оценку по дисциплине (модулю) и стимулирует углубленное изучение материала.

Самостоятельная работа – работа студентов, направленная на углубленное изучение отдельных тем и вопросов учебной дисциплины (модуля).

В процессе самостоятельной работы студенты взаимодействуют с рекомендованными материалами при минимальном участии преподавателя. Задачи студента включают работу с конспектами лекций (обработка текста), повторное изучение учебных материалов планов и тезисов ответов, изучение дополнительных тем, выполнение учебно-исследовательских заданий и другое.

Система оценивания результатов обучения по дисциплине (модулю)

Критерии получения уровня и оценивания сформированности компетенций по дисциплине (модулю) «Разметка наборов данных. Авито»

Оценивание уровня учебных достижений обучающихся по дисциплине (модулю) осуществляется в виде текущего контроля успеваемости и промежуточной аттестации.

Промежуточная аттестация по дисциплине (модулю) осуществляется в форме **экзамена**, при этом проводится оценка компетенций, сформированных по дисциплине.

Для оценивания текущего контроля успеваемости и промежуточной аттестации используется десятибалльная шкала оценивания, которая соотносится с традиционной пятибалльной шкалой следующим образом, если иное не указано в силлабусе дисциплины:

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
10	Отлично	Студент полностью владеет знаниями, изложенными в рабочей программе, и глубоко осмысляет дисциплину (модуль). Он самостоятельно и логически последовательно отвечает на все вопросы, акцентируя внимание на наиболее важном. Умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя ключевые моменты и устанавливая причинно-следственные связи. Четко формулирует ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты дисциплины (модуля) с практическими задачами.
9	Отлично	
8	Отлично	
7	Хорошо	Студент обладает знаниями предмета почти в полном объеме рабочей программы и самостоятельно, логически последовательно и всесторонне отвечает на все вопросы, акцентируя внимание на наиболее значимых моментах. Он умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя его ключевые аспекты и устанавливая причинно-следственные связи. Формулирует свои ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные ситуационные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты предмета с практическими задачами.
6	Хорошо	
5	Удовлетворительно	Студент обладает базовыми знаниями по дисциплине (модулю), но испытывает трудности при самостоятельных ответах и использует неточные формулировки. В ходе ответов он допускает ошибки, касающиеся сути вопросов. Студент способен решать только самые простые задачи и
4	Удовлетворительно	

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
		владеет лишь минимальным набором методов исследования.
3	Не сдан	Студент не овладел обязательным минимумом знаний по предмету и не может ответить на вопросы, даже если преподаватель задает дополнительные наводящие вопросы.
2	Не сдан	
1	Не сдан	

Дисциплина (модуль) «Разметка наборов данных. Авито» оценивается следующим образом:

Активность	Вес	Количество	Описание
Домашние задания	60%	10	Набор задач по темам недели
Экзамен	40%	1	Письменная или устная работа над заданием, направленным на проверку полученных знаний и навыков по дисциплине (модулю)

В рамках изучения дисциплины (модуля) возможно получение бонусных баллов.

Формула расчёта итоговой оценки по дисциплине (модулю) «Разметка наборов данных. Авито»: $0,6 \times \text{среднее за домашние задания} + 0,4 \times \text{экзамен}$.

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные домашние задания

Домашнее задание 1. Настройка краудсорсингового проекта

Спроектируйте задачу для разметки изображений товаров на платформе краудсорсинга. Разработайте подробную инструкцию для исполнителей, включающую примеры корректных и некорректных ответов. Создайте набор контрольных заданий (обучающих и проверочных) для отсева некачественных исполнителей. Рассчитайте бюджет проекта исходя из необходимого объема разметки и рыночной стоимости часа исполнителя.

Домашнее задание 2. Оптимизация разметки через активное обучение

Реализуйте пайплайн активного обучения для задачи классификации текстов объявлений. Обучите начальную модель на малой размеченной выборке. Используйте функцию неопределенности (entropy или margin) для отбора наиболее информативных объектов из неразмеченного пула. Сравните качество модели, обученной на случайно выбранной выборке и на выборке через активное обучение, при одинаковом бюджете разметки.

Домашнее задание 3. Сравнение разметки LLM и краудсорсинга

Выберите задачу классификации тональности отзывов. Выполните разметку тестовой выборки с помощью большой языковой модели (через промпты) и привлечением краудсорсинговых исполнителей. Проведите перекрестную валидацию результатов. Оцените стоимость, скорость и согласованность (Inter-Annotator Agreement) обоих методов. Сформулируйте рекомендацию по выбору инструмента для промышленного внедрения.

Задания для промежуточной аттестации по дисциплине (модулю)

№ п/п	Задание	Ответ	Компетенция
1.	Какой тип краудсорсинга предполагает участие экспертов?	Экспертная разметка	ОПК-1
2.	Что является основным преимуществом краудсорсинга перед штатными annotators?	Масштабируемость	ПК-3
3.	Как называется мера согласованности между разметчиками?	Коэффициент Каппы	ПК-3
4.	Какой метод отбора данных снижает затраты на разметку?	Активное обучение	ОПК-1
5.	Что используется для контроля качества в краудсорсинге?	Золотой стандарт	ПК-3
6.	Какая стратегия использует неразмеченные данные для улучшения модели?	Полуавтоматическое обучение	ОПК-1
7.	Как называется процесс создания новых данных из существующих?	Аугментация данных	ПК-3
8.	Какая модель генерирует изображения из шума?	Генеративно-состязательная сеть	ОПК-1
9.	Что означает аббревиатура SFT в контексте LLM?	Контролируемая донастройка	ПК-3
10.	Какой метод обучения использует человеческую обратную связь?	RLHF	ПК-3
11.	Что критично для инструкции краудсорсинга?	Однозначность формулировок	ОПК-1
12.	Как оценивается стоимость разметки за единицу?	Цена за задание	ПК-3
13.	Какой риск возникает при разметке силами пользователей?	Смещение выборки	УК-6
14.	Что гарантирует репрезентативность датасета?	Стратификация данных	ОПК-1
15.	Какая проблема характерна для генеративных данных?	Артефакты синтеза	ПК-3
16.	Как защищаются персональные данные при разметке?	Обезличивание	УК-6
17.	Что является основным критерием выбора между LLM и людьми?	Соотношение цена-качество	ПК-3
18.	Какой этический принцип нарушается при скрытой разметке?	Прозрачность сбора данных	УК-6
19.	Что измеряет кривая обучения модели?	Зависимость качества от объема данных	ОПК-1
20.	Кто несет ответственность за biases в размеченных данных?	Владелец датасета	УК-6