Федеральное государственное бюджетное образовательное

учреждение высшего образования

Московский государственный университет имени М.В. Ломоносова

|  |
| --- |
| **УТВЕРЖДАЮ****Ректор МГУ имени М.В. Ломоносова****В.А. Садовничий****«\_\_» \_\_\_\_\_\_\_\_ 2024 г.** |

**РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)**

**Наименование дисциплины (модуля):**

**Межфакультетский курс на английском языке**

**«Введение в обработку естественного языка и анализ текстов»**

*наименование дисциплины (модуля)*

**Уровень высшего образования:**

***Бакалавриат, магистратура, специалитет***

*бакалавриат, магистратура, специалитет*

**Направление подготовки / специальность:**

**По всем направлениям подготовки / специальностям**

*(код и название направления/специальности)*

**Направленность (профиль)/специализация ОПОП:**

**По всем направленностям (профилям)/специальностям ОПОП**

 *(если дисциплина реализуется в рамках направленности (профиля))*

Форма обучения:

Очная

*очная, очно-заочная*

Рабочая программа рассмотрена и одобрена

*Методической комиссией МГУ, на заседании Ученого совета МГУ*

\_\_.\_\_.2024

Москва 2024

Рабочая программа дисциплины (модуля) разработана в соответствии с самостоятельно установленными МГУ образовательными стандартами (ОС МГУ) для реализуемых основных профессиональных образовательных программ высшего образования (*программы бакалавриата, магистратуры, реализуемых последовательно по схеме интегрированной подготовки; программы специалитета; программы магистратуры)*

**1.** Место дисциплины (модуля) в структуре ОПОП ВО:

*Дисциплина «Введение в обработку естественного языка и анализ текстов» является межфакультетским курсом. Межфакультетские учебные курсы относятся к вариативной части и являются дисциплиной по выбору (элективной).*

**2.** Цели и задачи дисциплины (модуля):

Цель дисциплины — развить профессиональные информационно-коммуникационные и исследовательские компетенции студентов с возможностями применения приобретенных знаний, умений и навыков в прикладной области автоматизированной обработки научных текстов для своего исследования.

Задачи дисциплины:

— освоение основных понятий и терминов по дисциплинам корпусной и компьютерной лингвистики;

— получение опыта работы с различными подкорпусами текстов, относящихся к онлайн коммуникации;

— ознакомление с ключевыми методами обработки и анализа текстовых данных большими языковыми моделями.

**3.** Входные требования для освоения дисциплины (модуля), предварительные условия:

Перед началом освоения дисциплины «Основы обработки естественного языка технологиями искусственного интеллекта» студент должен изучить базовый курс «Информатика».

**4.** Результаты обучения по дисциплине (модулю):

|  |  |  |
| --- | --- | --- |
| **Компетенции выпускников** **(коды и формулировки)** | **Планируемые результаты обучения по дисциплине (модулю), сопряженные с компетенциями** | **Индикаторы (показатели) достижения компетенций** |
| **УК-10**. Способен использовать современные информационно-коммуникационные технологии в академической и профессиональной сферах. | Проектирование авторского подкорпуса  | **Знает:** **З1** - основные термины и понятия корпусной лингвистики;**З2** - основы кода на языке Python. Умеет:**У1** - составлять лингвистический подкорпус;**У2** - совершать очистку текстовых данных для автоматической обработки и анализа.**Владеет навыками:****В1** - отбора текстовых документов по критериям;**В2** - выбора библиотек для обработки естественного языка на всех этапах от сбора и подготовки данных к их анализу.**Демонстрирует готовность:****Г1** - собирать текстовые данные в цифровом пространстве;Г2 - преобразовывать текстовые данные. |
| **ОПК-6**. Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности. | Защита проекта | **Знает:** **З1** - основные термины и понятия компьютерной лингвистики;**З2** - основные термины и понятия раздела лингвистики для междисциплинарного направления искусственного интеллекта.**Умеет:****У1** - подбирать по критериям инструменты на базе нейронных сетей;**У2** - совершать автоматический анализ данных.**Владеет навыками:****В1** - фильтрации функций нейронных сетей для обработки текста;**В2** - визуализации результатов автоматического лингвистического анализа.**Демонстрирует готовность:****Г1** - совершать автоматический лингвистический анализ;Г2 - визуализировать результаты. |

**5.** Объем дисциплины

Общая трудоемкость дисциплины составляет 1 з.е. (48 часов), включая 24 часа на занятия лекционного типа и 24 часов на самостоятельную работу обучающихся).

|  |  |
| --- | --- |
| **Вид учебной работы** | **Всего часов** |
| **Аудиторные занятия (всего)** | **24** |
| В том числе: |  |
| Лекции (Л) | 24 |
| Практические занятия (ПЗ) | - |
| Семинары (С) | - |
| Лабораторные работы (ЛР) | - |
| **Самостоятельная работа (всего)** | **12** |
| В том числе: |  |
| Подготовка к семинарским занятиям | - |
| Подготовка домашних заданий | 11 |
| Решение профессиональных задач и т.д. | - |
| Подготовка к аттестации  | 1 |
| Вид промежуточной аттестации (зачет, экзамен) | зачет |
| Общая трудоемкость: часызачетные единицы | **36** |
| **1** |

**6.** Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий

**Содержательный план дисциплины**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **№ п/п** | **Наименование раздела дисциплины** **(элемента модуля) и его содержание** | **Форма текущего контроля успеваемости *(наименование)*** | **Лекции** | **Практические** **занятия**  | **Семинары**  | **Самостоятельная работа**  | **Всего** |
| 1.  | **Тема 1. Введение в проблематику ИИ как междисциплинарного направления: связь с когнитивной лингвистикой**Термин «искусственный интеллект» и неточности его перевода; история развития бенчмарков (Тест Тьюринга и «Китайская комната»); ограниченный и сильный искусственный интеллект (Artificial General Intelligence, AGI); наука о больших данных (особенности машинное и глубокое обучение нейронных сетей при работе с текстовыми данными). | Мобильный квест на сравнение помощников на основе GPT-4o и Алисы от Яндекс по предметным областям. | **2** |  |  | **1** | **3** |
| 2. | **Тема 2. Основы дистрибутивной семантики и введение в обработку естественного языка (Natural Language Processing, NLP).** Дистрибутивная гипотеза, закон Ципфа и представление языковых единиц в многомерном пространстве (метод векторизации); обработка естественного языка (история, принципы и сферы применения NLP). | Веб-квест по изучению прикладных сфер применения NLP в предметных областях.  | **2** |  |  | **1** | **3** |
| 3. | **Тема 3. Большие языковые модели и технологии генеративного ИИ.** Типология, архитектура и методы работы больших языковых моделей; эмбеддинги; обучение с учителем и без учителя, обучение с подкреплением; continuous bag of words vs skip-gram; word matrix, механизм внимания; задачи классификации и распознавания именных сущностей. | Проблемно-поисковое задание на ресурсе SkeLL для иллюстрации эмбеддингов.  | **2** |  |  | **1** | **3** |
| 4. | **Тема 4. Информационный поиск в эпоху генеративного искусственного интеллекта.** Принципы формирования баз данных научных статей; индексы цитирования; возможности и ограничения нейронных сетей для теоретического обзора. | Веб-квест насравнение Google Scholar с Semantic Scholar, Consensus, etc.  | **2** |  |  | **1** | **3** |
| 5. | **Тема 5. Машинный перевод по предметным областям.** История машинного перевода; системы онлайн перевода и профессиональные автоматические системы перевода; ознакомление с системами устного синхронного перевода. | Задание по пред- и постредактированию текстов по предметным областям для машинного перевода в DeepL и SmartCat. | **2** |  |  | **1** | **3** |
| 6. | **Тема 6. Основы компьютерной лингвистики и морфологический анализ.** Лексическая неоднозначность как основная проблема разметки текстовых данных (омонимия, полисемия, конверсия и т.д.); токенизация на уровне слова; предиктивный морфологический анализ, методы лемматизации (для русского языка на основе теории Щербы Л.В., грамматического словаря Зализняка А.А. и ПО MyStem Ильи Сегаловича) и стеминга (для английского языка методом обработки суффиксов без словаря основ на основе ПО Snowball); частеречная разметка (POS). | Проектирование авторского подкорпуса текстов научных статей по тематике исследования (планируемый результат: база данных). | **2** |  |  | **1** | **3** |
| 7. | **Тема 7. Основы программирования на Python для обработки естественного языка.** Блокнот как среда обработки (Integrated Development Environment, IDE); литералы и функции строк; операции и методы строк; f строки; кортежи и словари; функции if/else, синтаксис и принципы работы чат-бота. | Практикум по python в Google Colab. | **2** |  |  | **1** | **3** |
| 8. | **Тема 8. Основы работы с открытыми библиотеками обработки естественного языка.** Библиотеки numpy & pandas для обработки данных; фреймворки для разработки (keras vs tensorflow vs pytorch); этапы подготовки текстовых данных (очистка и преобразование); встроенные библиотеки python и модули NLTK для токенизации, удаления знаков препинания и стоп-слов, подсчета частотности и т.д. | Подготовка авторского подкорпуса текстов для анализа в Google Colab (планируемый результат: список частотных языковых единиц по теме). | **2** |  |  | **1** | **3** |
| 9. | **Тема 9. Основы корпусной лингвистики для анализа авторского подкорпуса.** История развития лингвистических корпусов; национальные корпусы; Национальный корпус русского языка (НКРЯ); параллельные корпусы; функции корпусного менеджера; конкорданс; N-граммы; лексический профиль слова; дисперсия и облака слов; методы структурного лингвистического анализа (дистрибутивный и диахронический анализ); коэффициенты оценки коллокаций LogDice, Loglikelihood, MI3, T-score, etc. | «Портрет слова» для терминов и персоналий по предметным областям (планируемый результат: статистические данные по употреблению выделенных языковых единиц).  | **2** |  |  | **1** | **3** |
| 10. | **Тема 10. Обработка и визуализация результатов лингвистического анализа.** Принципы визуализации лингвистического анализа; способы визуализации числового вектора: облако слов, ментальная карта и т.д.; понятие графа. | Граф как презентация результатов лингвистического анализа авторского подкорпуса текстов научных статей по тематике исследований. | **2** |  |  | **1** | **3** |
| 11. | **Тема 11. Основы стилометрии в аспекте авторства научных работ.** История развития стилометрии; Дельта Барроуза; проблема антиплагиата; современные инструменты выявления плагиата, в том числе сгенерированного искусственным интеллектом. | Определение авторства научных материалов в R-Stylo.  | **2** |  |  | **1** | **3** |
| 12. | **Тема 12. Чат-бот для сбора количественных и качественных данных по исследованию.** Методы опроса, фокус-группы и интервьюирования; принципы распознавания речи; возможности автоматизирования количественных и качественных данных для исследования. | Практическое задание «Маршрут бота» для исследования. | **2** |  |  | **1** | **3** |
|  | **Итого по курсу:** | **24** |  |  | **12** | **36** |

**7.** Фонд оценочных средств (ФОС, оценочные и методические материалы) для оценивания результатов обучения по дисциплине.

**7.1.** Критерии оценивания

12 тестов по теории (10 вопросов, 15-20 минут на выполнение): 5 баллов \* 12 тестов = 60 баллов (максимально);

12 заданий (30-40 минут на выполнение, интерактивные и практические задания; практические задания составляют подготовку проекта «Корпусное исследование статей по проблематике научного интереса студента»): 10 баллов \* 12 заданий = 120 баллов (максимально);

защита проекта – 20 баллов (максимально).

Промежуточная аттестация методом портфолио (зачет) проводится на последнем занятии.

Итого: 200 баллов (максимально); зачет: выше 150 баллов.

**7.2.** Список вопросов к зачету:

1. 1. ИИ как междисциплинарное направление: связь с когнитивной лингвистикой.
2. Основы дистрибутивной семантики и введение в обработку естественного языка.
3. Большие языковые модели и технологии генеративного ИИ.
4. Информационный поиск в эпоху генеративного искусственного интеллекта.
5. Машинный перевод по предметным областям.
6. Основы компьютерной лингвистики и морфологический анализ.
7. Основы работы с открытыми библиотеками для обработки естественного языка.
8. Методы лингвистического анализа авторского подкорпуса.
9. Обработка и визуализация результатов лингвистического анализа.
10. Основы стилометрии в аспекте авторства научных работ.

**8.** Ресурсное обеспечение.

**8.1.** Перечень основной и дополнительной учебной литературы

Основная литература

1. Авраменко А.П. Большие языковые модели в лингвистике и лингводидактике. 2 изд. М., 2024.

Дополнительная литература

1. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных. М., 2017.
2. Barthélemy F., et alt. Natural language processing for public services, EU, 2022.
3. Chandler D. Semiotics: the basics, 4th ed., Routledge, 2022.
4. Jurafsky D., et alt. Speech and Language Processing, 3d ed., Stanford, 2024.
5. Lenci A., et alt. Distributional Semantics, Cambridge University Press, 2023.
6. Manning Ch., et alt. Introduction to Information Retrieval, Cambridge University Press. 2008.

**8.2.** Перечень лицензионного программного обеспечения

* Google Colab
* SketchEngine
* Consensus
* AntCont
* R-Stylo

**8.3.** Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

* НКРЯ
* GitHub

**8.4.** Описание материально-технической базы

Ноутбук, мультимедийный проектор, акустические колонки, экран, указка/лазерная указка.

**9.** Язык преподавания

Английский язык.

**10.** Преподаватели

Авраменко Анна Петровна – к.п.н., доцент кафедры лингвистики и информационных технологий ФИЯР МГУ имени М.В. Ломоносова.

**11.** Авторы программы

Авраменко Анна Петровна – к.п.н., доцент кафедры лингвистики и информационных технологий ФИЯР МГУ имени М.В. Ломоносова.