Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский ядерный университет «МИФИ»

ИНСТИТУТ ФУНДАМЕНТАЛЬНЫХ ПРОБЛЕМ СОЦИАЛЬНО-ГУМАНИТАРНЫХ НАУК

КАФЕДРА ПСИХОЛОГИИ, СОЦИОЛОГИИ И АНТРОПОЛОГИИ

ОДОБРЕНО УМС ИИКС

Протокол № 8/1/2025

от 25.08.2025 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

АВТОМАТИЧЕСКАЯ ОБРАБОТКА ТЕКСТОВ В КОРПУСНОЙ ЛИНГВИСТИКЕ

Направление подготовки (специальность) [1] 09.03.04 Программная инженерия

Семестр	**Трудоемкость, кред.**	**Общий объем курса, час.**	**Лекции, час.**	**Практич. занятия, час.**	**Лаборат. работы, час.**	**В форме практической подготовки, час.**	**СРС, час.**	**КСР, час.**	**Форма(ы) контроля, экз./зач./КР/КП**
4	1	36	24	0	0		12	0	З
Итого	1	36	24	0	0	0	12	0	

АННОТАЦИЯ

Курс автоматической обработки текстов в корпусной лингвистике знакомит студентов с проблематикой задач обработки текстов на естественном языке (АОТ; автоматической обработки текста) и основными способами работы с цифровыми инструментами, применимыми в гуманитарных науках. Студенты познакомятся с основными методами, подходами и методами такой обработки. В частности, студенты изучат основные определения корпусной и компьютерной лингвистики, познакомятся с моделями тематического моделирования, векторной семантики и другими разделами компьютерной лингвистики, разметки текстов синтаксической, морфологической и другими видами аннотирования. Знакомство с набором инструментов цифровой гуманитаристики (digital humanities) позволит слушателям создавать новые самостоятельные исследовательские проекты, обрабатывать большие объемы информации, делать гуманитарные исследования более репрезентативными и актуальными. Познакомившись с «палитрой» методов DH — от оцифровки и семантического анализа до добычи и визуализации данных и работы с геоинформационными системами — слушатели смогут применить их на практике. Благодаря освоенному материалу слушатели смогут значительно обогатить методологическую базу и поймут, как ее применять к широкому спектру гуманитарных исследовательских проблем. Логика курса направлена на практическое освоение цифровых методов и позволяет применять их на практике.

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ УЧЕБНОЙ ДИСЦИПЛИНЫ

Цель раздела, посвященного автоматической обработки текстов в корпусной лингвистике - познакомить студентов с задачами на стыке двух дисциплин: программирования и лингвистики. Курс научит рассматривать лингвистические вопросы в разрезе компьютерных методов и ставить перед собой и решать такие лингвистические задачи. Студенты познакомятся с лингвистическими процессорами и различными моделями в компьютерной лингвистике.

Задачи дисциплины заключаются в развитии следующих знаний, умений и навыков:

• владение основной терминологией корпусной и компьютерной лингвистики;

• знание основных направлений обработки текстов на естественном языке;

• владение инструментарием обработки;

• умение корректно оценить результат такой обработки.

Цели раздела, посвященного цифровым методам в гуманитарных науках

• изучить различные теоретические и прикладные аспекты цифровых гуманитарных наук;

• познакомить с широким спектром цифровых гуманитарных проектов, а также методов и инструментов для их реализации;

• познакомить с современным исследовательским полем DH, дать инструментарий для анализа и критики исследований и проектов в DH;

• сформировать навыки планирования, разработки и оценки цифровых гуманитарных проектов.

В результате освоения студенты овладеют навыками:

• критически обсуждать цифровые гуманитарные науки в свете текущих теорий и методов;

• объяснять, как цифровые гуманитарные методы и практики применяются в различных дисциплинах;

• выдвигать гипотезу и ставить исследовательский вопрос в рамках исследования в области цифровых гуманитарных наук.

2. МЕСТО УЧЕБНОЙ ДИСЦИПЛИНЫ В СТРУКТУРЕ ООП ВО

Данная учебная дисциплина представляет собой один из курсов блока элективных дисциплин и имеет междисциплинарный характер. Курс связан как с другими лингвистическими дисциплинами, так и непосредственно с компьютерными. Курс тесно связан с блоком дисциплин, связанных с введением в языкознание и требует общего понимания основной лингвистической терминологии. Студентам также понадобятся навыки программирования и умение писать алгоритмы для того, чтобы использовать некоторые модели АОТ.3.

3. ФОРМИРУЕМЫЕ КОМПЕТЕНЦИИ И ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ

Универсальные и(или) общепрофессиональные компетенции:

Код и наименование компетенции	**Код и наименование индикатора достижения компетенции**
УК-1 [1] – Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	З-УК-1 [1] – Знать: методики сбора и обработки информации; актуальные российские и зарубежные источники информации в сфере профессиональной деятельности; метод системного анализа У-УК-1 [1] – Уметь: применять методики поиска, сбора и обработки информации; осуществлять критический анализ и синтез информации, полученной из разных источников В-УК-1 [1] – Владеть: методами поиска, сбора и обработки, критического анализа и синтеза информации; методикой системного подхода для решения поставленных задач
УК-4 [1] – Способен осуществлять деловую коммуникацию в устной и письменной формах на государственном языке Российской Федерации и иностранном(ых) языке(ах)	З-УК-4 [1] – Знать: принципы построения устного и письменного высказывания на русском и иностранном языках; правила и закономерности деловой устной и письменной коммуникации У-УК-4 [1] – Уметь: применять на практике деловую коммуникацию в устной и письменной формах, методы и навыки делового общения на русском и иностранном языках; методикой составления суждения в межличностном деловом общении на русском и иностранном языках В-УК-4 [1] – Владеть: навыками чтения и перевода текстов на иностранном языке в профессиональном общении; навыками деловых коммуникаций в устной и письменной

	форме на русском и иностранных языках; методикой составления суждения в межличностном деловом общении на русском и иностранном языках

4. ВОСПИТАТЕЛЬНЫЙ ПОТЕНЦИАЛ ДИСЦИПЛИНЫ

Направления/цели воспитания	**Задачи воспитания (код)**
Профессиональное воспитание	Создание условий, обеспечивающих, формирование научного мировоззрения, культуры поиска нестандартных научно-технических/практических решений, критического отношения к исследованиям лженаучного толка (В19)

5. СТРУКТУРА И СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Разделы учебной дисциплины, их объем, сроки изучения и формы контроля:

№ п.п	**Наименование раздела учебной дисциплины**	**Недели**	**Лекции/ Практ. (семинары)/ Лабораторные работы, час.**	**Обязат. текущий контроль (форма*, неделя)**	**Максимальный балл за раздел****	**Аттестация раздела (форма*, неделя)**	**Индикаторы освоения компетенции**
	4 Семестр						
1	Знакомство с основными определениями автоматической обработки текстов в корпусной лингвистике	1-2	4/0/0		8	КИ-2	З-УК-1, У-УК-1, В-УК-1, З-УК-4, У-УК-4, В-УК-4
2	Направления в автоматической обработке текстов на естественном языке	3-6	8/0/0		16	КИ-6	З-УК-1, У-УК-1, В-УК-1, З-УК-4, У-УК-4, В-УК-4
3	Знакомство с основными определениями цифровой гуманитаристики	7-7	2/0/0		6	КИ-7	З-УК-1, У-УК-1, В-УК-1, З-УК-4, У-УК-4, В-УК-4
4	Направления в	8-12	10/0/0		20	КИ-12	З-УК-1,

	цифровой гуманитаристике						У-УК-1, В-УК-1, З-УК-4, У-УК-4, В-УК-4
	Итого за 4 Семестр		24/0/0		50		
	Контрольные мероприятия за 4 Семестр				50	З	З-УК-1, У-УК-1, В-УК-1, З-УК-4, У-УК-4, В-УК-4

* – сокращенное наименование формы контроля

** – сумма максимальных баллов должна быть равна 100 за семестр, включая зачет и (или) экзамен

Сокращение наименований форм текущего контроля и аттестации разделов:

Обозначение	**Полное наименование**
КИ	Контроль по итогам
З	Зачет

КАЛЕНДАРНЫЙ ПЛАН

Недели	**Темы занятий / Содержание**	**Лек., час.**	**Пр./сем., час.**	**Лаб., час.**
	4 Семестр	24	0	0
1-2	**Знакомство с основными определениями автоматической обработки текстов в корпусной лингвистике**	4	0	0
1	**Общее знакомство с терминологией компьютерной лингвистики** Что такое корпус. Задачи, решаемые с помощью корпуса. Корпус как датасет. Свойства корпуса. Токен, токенизация. Нормализация текста, аннотация. Предобработка корпуса. Регулярные выражения.	Всего аудиторных часов		
		2	0	0
		Онлайн		
		0	0	0
2	**Статистическая обработка корпуса** Статистический анализ текста по лексическим параметрам. Синтаксическая разметка, построение деревьев, сбор статистики по ним. Ветвление, вложение групп.	Всего аудиторных часов		
		2	0	0
		Онлайн		
		0	0	0
3-6	**Направления в автоматической обработке текстов на естественном языке**	8	0	0
3	**Семантическая разметка** Простая семантика: представление текста как вектора (мешок слов, tf-idf), векторная семантика.	Всего аудиторных часов		
		2	0	0
		Онлайн		
		0	0	0
4	**Компьютерная семантика: моделирование и анализ тональности** Определение темы. Модели тематического	Всего аудиторных часов		
		2	0	0
		Онлайн		

	моделирования: LDA, PLSA. Понятие анализа тональности. Основные модели анализа.	0	0	0
5	**Извлечение информации** Извлечение отношений. Извлечение сущностей, событий, построение саммари.	Всего аудиторных часов		
		2	0	0
		Онлайн		
		0	0	0
6	**Генерация текстов, машинный перевод** Задача машинного перевода и общие определения. Задача генерации текстов. Простые модели генерации текста, марковские цепи.	Всего аудиторных часов		
		2	0	0
		Онлайн		
		0	0	0
7-7	**Знакомство с основными определениями цифровой гуманитаристики**	2	0	0
7	**История DH. Данные в цифровых гуманитарных исследованиях** Периоды становления DH: 1950е-1970е, 1980е-1990е, 2000е-н.в. Гуманитарные данные: Data vs Capta; Big data vs Small data vs Smart data. Этика данных. Критический подход к данным. Где брать данные для своего исследования и как ими пользоваться. Создание метаданных, или Аннотирование предметов культурного наследия. Формат TEI.	Всего аудиторных часов		
		2	0	0
		Онлайн		
		0	0	0
8-12	**Направления в цифровой гуманитаристике**	10	0	0
8	**Культуромика: исследование культуры и языка с помощью больших текстовых данных** Культуромика как научное направление. Google Books как исследовательская база. Возможности поиска с помощью Google Ngram Viewer. Дистрибутивные языковые модели и историческая семантика.	Всего аудиторных часов		
		2	0	0
		Онлайн		
		0	0	0
9	**Базы данных в гуманитарных науках: модели, структуры, связанные данные** Цифровой поворот и структурирование информации в гуманитарных науках. Специфика баз данных в гуманитарных науках. Типы данных и семантический разрыв. Подходы к созданию гуманитарно-ориентированных баз данных.	Всего аудиторных часов		
		2	0	0
		Онлайн		
		0	0	0
10	**Цифровое литературоведение** “Дальнее чтение”: принципы, направления, открытия. Цифровой анализ поэзии. Цифровой анализ прозы. Стилометрия и большие языковые модели.	Всего аудиторных часов		
		2	0	0
		Онлайн		
		0	0	0
11	**Распознавание и расшифровка письменных исторических документов** Методологические принципы и этапы ручной расшифровки. Проект «Прожито». Технологии computer-assisted расшифровки: инструмент Transcribus. Применение компьютерного зрения и глубинного обучения.	Всего аудиторных часов		
		2	0	0
		Онлайн		
		0	0	0
12	**Геоинформационные системы: подходы, методики, данные** Теоретические основы геоинформационных систем. Структура геоинформационного проекта и используемые данные. Основные этапы работы над ГИС-проектом Пространственный анализ средствами	Всего аудиторных часов		
		2	0	0
		Онлайн		
		0	0	0

	геоинформационных систем.			

Сокращенные наименования онлайн опций:

Обозначение	**Полное наименование**
ЭК	Электронный курс
ПМ	Полнотекстовый материал
ПЛ	Полнотекстовые лекции
ВМ	Видео-материалы
АМ	Аудио-материалы
Прз	Презентации
Т	Тесты
ЭСМ	Электронные справочные материалы
ИС	Интерактивный сайт

6. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

В процессе преподавания используются авторские, мультимедийные, диалоговые и тестовые технологии. Важное значение придается вовлечению студентов в практическую исследовательскую работу. Особое значение придаётся практической работе студентов по разметке и освоению основных методологий автоматического анализа текста на естественном языке.

7. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

Фонд оценочных средств по дисциплине обеспечивает проверку освоения планируемых результатов обучения (компетенций и их индикаторов) посредством мероприятий текущего, рубежного и промежуточного контроля по дисциплине.

Связь между формируемыми компетенциями и формами контроля их освоения представлена в следующей таблице:

Компетенция	**Индикаторы освоения**	**Аттестационное мероприятие (КП 1)**
УК-1	З-УК-1	З, КИ-2, КИ-6, КИ-7, КИ-12
	У-УК-1	З, КИ-2, КИ-6, КИ-7, КИ-12
	В-УК-1	З, КИ-2, КИ-6, КИ-7, КИ-12
УК-4	З-УК-4	З, КИ-2, КИ-6, КИ-7, КИ-12
	У-УК-4	З, КИ-2, КИ-6, КИ-7, КИ-12
	В-УК-4	З, КИ-2, КИ-6, КИ-7, КИ-12

Шкалы оценки образовательных достижений

Шкала каждого контрольного мероприятия лежит в пределах от 0 до установленного максимального балла включительно. Итоговая аттестация по дисциплине оценивается по 100-балльной шкале и представляет собой сумму баллов, заработанных студентом при выполнении заданий в рамках текущего и промежуточного контроля.

Итоговая оценка выставляется в соответствии со следующей шкалой:

Сумма баллов	Оценка по 4-х балльной шкале	Отметка о зачете	Оценка ECTS
90-100	5 – *«отлично»*	*«зачтено»*	A
85-89	4 – «*хорошо*»		B
75-84			C
70-74			D
65-69	3 – «*удовлетворительно*»		
60-64			E
ниже 60	2 – «*неудовлетворительно*»	*«не зачтено»*	F

Оценка «отлично» соответствует глубокому и прочному освоению материала программы обучающимся, который последовательно, четко и логически стройно излагает свои ответы, умеет тесно увязывать теорию с практикой, использует в ответах материалы монографической литературы.

Оценка «хорошо» соответствует твердым знаниям материала обучающимся, который грамотно и, по существу, излагает свои ответы, не допуская существенных неточностей.

Оценка «удовлетворительно» соответствует базовому уровню освоения материала обучающимся, при котором освоен основной материал, но не усвоены его детали, в ответах присутствуют неточности, недостаточно правильные формулировки, нарушения логической последовательности.

Отметка «зачтено» соответствует, как минимум, базовому уровню освоения материала программы, при котором обучающийся владеет необходимыми знаниями, умениями и навыками, умеет применять теоретические положения для решения типовых практических задач.

Оценку «неудовлетворительно» / отметку «не зачтено» получает обучающийся, который не знает значительной части материала программы, допускает в ответах существенные ошибки, не выполнил все обязательные задания, предусмотренные программой. Как правило, такие обучающиеся не могут продолжить обучение без дополнительных занятий.

8. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

ОСНОВНАЯ ЛИТЕРАТУРА:

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА:

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ:

Специальное программное обеспечение не требуется

LMS И ИНТЕРНЕТ-РЕСУРСЫ:

https://online.mephi.ru/

http://library.mephi.ru/

9. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Специальное материально-техническое обеспечение не требуется

10. УЧЕБНО-МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ДЛЯ СТУДЕНТОВ

Для эффективного усвоения курса необходимо регулярное посещение и активное участие студентов на занятии, выполнение всех практических заданий на семинарах и выполнение итогового задания. Дисциплина способствует формированию языковых и технических компетенций, пониманию особенностей междисциплинарности компьютерной лингвистики. Курс пригодится всем студентам, планирующим в дальнейшем работать с текстом как источником данных для анализа, машинного обучения и других подобных задач.

11. УЧЕБНО-МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ДЛЯ ПРЕПОДАВАТЕЛЕЙ

На лекциях студентам сообщаются новые сведения, систематизируется и обобщается накопленный запас знаний, формируются на этой основе познавательные и профессиональные интересы. Преподаватель, читая лекцию, должен стремиться увлечь студентов, активно воздействовать на их эмоции, вызвать интерес к учебному предмету, стремление постоянно пополнять знания.

Структура каждой лекции складывается из вступления, основной части и заключения. Во вступлении кратко формулируется тема, сообщается план и задачи, перечисляется литература к лекции, показывается связь с предшествующим материалом, характеризуется теоретическая и практическая значимость темы. В основной части всесторонне раскрывается содержание проблемы, обосновываются ключевые идеи и положения, осуществляется их конкретизация, показываются связи и отношения, дается оценка сложившейся практике и научным основаниям, раскрываются перспективы развития, формулируются промежуточные выводы. В заключительной части подводится итог, кратко повторяются и обобщаются основные положения, формулируются общие выводы, излагаются рекомендации по выполнению самостоятельной работы. Необходимо также предусмотреть время для ответов на возможные вопросы студентов.

Преподавательский конспект каждой лекции должен содержать: тему (точно сформулированную); цель и задачи; перечень вопросов, освещаемых на лекции; реферативное изложение содержания (с выделением основного и второстепенного).

При чтении лекции необходимо поддерживать высокий научный уровень излагаемой информации; обеспечивать доказательность и достоверность высказываемых суждений; ясно и точно излагать мысли и активизировать мышление слушателей; выделять интонационно каждый раздел; в каждом вопросе вычленять главное (для запоминания) и второстепенное (для иллюстрации); четко проговаривать термины, расшифровывать и записывать их на доске; обращать внимание студентов на наиболее значимые понятия и модели.

Самостоятельные занятия студентов по выполнению домашних заданий представляют собой логическое продолжение аудиторных занятий, проводятся по заданию преподавателя, который инструктирует студентов и устанавливает сроки выполнения задания. Их главное

назначение состоит в закреплении знаний и умений, полученных на аудиторных занятиях, отработке навыков, усвоении нового материала.

Автор(ы):

Климов Александр Антонович