

УТВЕРЖДАЮ
Проректор по учебной работе и
качеству образования

_____ И. А. Долгова

30 сентября 2026

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Б1.О.11

**Искусственный интеллект в
лингвистике**

Направление подготовки:	45.04.04 Интеллектуальные системы в гуманитарной среде
Профиль подготовки:	Цифровая лингвистика
Квалификация:	магистр
Форма обучения:	очная
Год начала подготовки:	2027

Рабочая программа составлена в соответствии с:

- федеральным государственным образовательным стандартом высшего образования – магистратура по направлению подготовки 45.04.04 Интеллектуальные системы в гуманитарной среде, утверждённым приказом Министерства науки и высшего образования Российской Федерации от 13 августа 2020 года № 1003.

Разработчики программы: Макаров А. А., доктор технических наук, профессор; Кириллов А. Г., кандидат филологических наук, доцент

Рабочая программа согласована с руководителем образовательной программы 45.04.04 Интеллектуальные системы в гуманитарной среде. Рабочая программа рассмотрена и одобрена на заседании кафедры информационных систем и компьютерных технологий 10 сентября 2026, протокол № 2.

1. ОБЩАЯ ХАРАКТЕРИСТИКА

1.1. Цель и задачи дисциплины

Цель изучения дисциплины — формирование у обучающихся системы фундаментальных знаний, умений и практических навыков в области применения технологий искусственного интеллекта, машинного обучения и нейросетевого моделирования для автоматизированного анализа текстов и решения прикладных лингвистических задач.

Задачи изучения дисциплины:

- изучение теоретических основ машинного обучения (Machine Learning), архитектур искусственных нейронных сетей и современных подходов к обработке естественного языка (Natural Language Processing — NLP);
- формирование навыков проектирования программных конвейеров (пайплайнов) автоматизированного интеллектуального анализа текстовых массивов, включая векторизацию, классификацию и извлечение знаний;
- освоение принципов работы, возможностей и ограничений современных контекстуальных моделей и больших языковых моделей (LLM), а также методологии промт-инжиниринга;
- Развитие умений математически и программно оценивать качество, валидность и эффективность обучения нейросетевых моделей при работе с разнородными языковыми данными.

1.2. Требования подготовленности обучающегося к освоению содержания дисциплины

Освоение дисциплины «Искусственный интеллект в лингвистике» опирается на прочный ИТ- и математический фундамент, заложенный на предыдущих этапах обучения в 1 и 2 семестрах (дисциплины «Математика в гуманитарных науках», «Язык программирования Python», «Корпусная лингвистика», «Методика лингвистического исследования»).

Обучающийся перед началом освоения дисциплины должен:

Знать:

- базовые понятия математического анализа, элементы математической логики и теории вероятностей;
- синтаксис языка Python, принципы объектно-ориентированного программирования и стандартные библиотеки обработки данных;
- принципы организации, структуры и разметки лингвистических корпусов.

Уметь:

- разрабатывать прикладные программные скрипты на Python для автоматизированного сбора (парсинга) и предобработки текстовых массивов;
- использовать инструменты корпусного анализа для извлечения частотных характеристик текста;
- работать со специализированными средами разработки (Jupyter Notebook / VS Code) и системами контроля версий Git.

Владеть:

- навыками программирования и отладки программного кода;
- методами базовой математической и статистической обработки текстовой информации.

1.3. Планируемые результаты обучения по дисциплине

Код и наименование компетенции	Код и наименование индикатора достижения компетенции	Результаты обучения по дисциплине
ОПК-5. Способен применять новые информационные технологии в гуманитарных областях знаний с использованием средств интеллектуального анализа данных, машинного обучения, компьютерной лингвистики и представления знаний	ОПК-5.1. Выбирает методы интеллектуального анализа данных, машинного обучения и компьютерной лингвистики для автоматизированной обработки текстовой информации	ОПК-5.1.2. Знает теоретические основы машинного обучения, принципы работы нейросетевых архитектур (включая трансформеры и LLM) и методы извлечения знаний из неструктурированного текста
	ОПК-5.2. Реализует алгоритмы предварительной обработки, анализа и представления знаний на основе больших массивов языковых данных	ОПК-5.2.2. Умеет проектировать пайплайны обработки текста (векторизация, эмбединги, классификация) и настраивать (Fine-Tuning/Prompt Engineering) модели ИИ для лингвистических задач
	ОПК-5.3. Применяет специализированные программные библиотеки, инструменты ИИ и технологии интеллектуального анализа для решения прикладных лингвистических задач	ОПК-5.3.2. Владеет навыками применения алгоритмов машинного обучения (токенизация, сентимент-анализ, NER) с использованием профильных ИТ-инструментов и ИИ-платформ

2. ОБЪЁМ, СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

2.1. Объём дисциплины

Семестр	Контактная работа			СР	Форма ПА	Итоговый объём, часов/з. е.
	Л	П	ПА			
3	–	32	–	76	зачёт	108/3
Итого:	–	32	–	76		108/3

Л — лекции, П — все виды занятий семинарского типа, ПА — промежуточная аттестация, СР — самостоятельная работа обучающегося.

2.2. Структура дисциплины

Наименование тематического раздела дисциплины	Количество часов		
	Л	П	СР
Введение в интеллектуальный анализ данных и инженерии знаний		4	10
Архитектура информационного поиска и предобработка текстовых данных		6	14
Методология и алгоритмы машинного обучения в лингвистике		6	14
Векторные представления слов: статические и контекстуальные эмбединги		8	18
Большие языковые модели (LLM) и генеративный искусственный интеллект		8	20
Всего:	–	32	76

2.3. Содержание тематических разделов дисциплины

Раздел 1. Введение в интеллектуальный анализ данных и инженерии знаний
Архитектура интеллектуальных систем в гуманитарной сфере. Данные, информация и знания. Классические модели представления знаний (логические, продукционные, фреймовые, семантические сети) и особенности их применения к естественному языку. Понятие онтологии и графов знаний. Направления, задачи и специфика компьютерной лингвистики на современном этапе. Распознавание и синтез речи (технологии ASR и TTS).
Раздел 2. Архитектура информационного поиска и предобработка текстовых данных
Модели информационного поиска (векторная, вероятностная, булева). Индексация текстовых массивов. Полный цикл предобработки неструктурированных языковых данных (Data Preprocessing Pipeline): токенизация, удаление стоп-слов, стемминг, лемматизация, морфологический и синтаксический парсинг текста. Прикладные библиотеки Python для предобработки (NLTK, SpaCy).
Раздел 3. Методология и алгоритмы машинного обучения в лингвистике
Понятие машинного обучения (Machine Learning). Обучение с учителем (Supervised Learning), обучение без учителя (Unsupervised Learning) и обучение с подкреплением (Reinforcement

Learning). Количественные и статистические показатели качества обучения (Accuracy, Precision, Recall, F-мера, ROC-AUC). Применение классических алгоритмов (Наивный Байес, Метод опорных векторов, Логистическая регрессия) для автоматической классификации и кластеризации текстов, анализа тональности (Sentiment Analysis).

Раздел 4. Векторные представления слов: статические и контекстуальные эмбединги

Проблема дистрибутивной семантики. Переход от разреженных векторов (One-Hot Encoding, TF-IDF) к плотным векторным представлениям. Статические модели эмбедингов (Word2Vec, FastText, GloVe): архитектуры Continuous Bag-of-Words (CBOW) и Skip-gram. Ограничения статических моделей. Контекстуальные эмбединги (ELMo, BERT). Архитектура Transformer (механизм внимания — Attention). Применение предобученных моделей для решения задач компьютерной лингвистики (NER, извлечение отношений).

Раздел 5. Большие языковые модели (LLM) и генеративный искусственный интеллект

Понятие и эволюция больших языковых моделей (классы моделей GPT, LLaMA, YandexGPT и др.). Возможности, технологические ограничения и проблема «галлюцинаций» больших языковых моделей. Методология проектирования эффективных запросов (Prompt Engineering). Системы и методы верификации, автоматической валидации и фактчекинга результатов работы LLM. Сценарии интеграции генеративных нейросетей в профессиональную деятельность цифрового лингвиста.

2.4. Организация учебных занятий

Дисциплина может реализовываться с применением электронного обучения и дистанционных образовательных технологий.

3. ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

3.1. Учебная литература

3.1.1. Основная

1. Иванченко, Д. А. Нейросетевые технологии в образовании : возможности и применение : методическое пособие : [16+] / Д. А. Иванченко. – Москва : Директ-Медиа, 2025. – 88 с. : ил., табл. – Режим доступа: по подписке. – URL: <https://biblioclub.ru/index.php?page=book&id=718842>.
2. Кревецкий, А. В. Основы технологий искусственного интеллекта : учебное пособие : [16+] / А. В. Кревецкий, Ю. А. Ипатов, Н. И. Роженцова ; под общ. ред. А. В. Кревецкого ; Поволжский государственный технологический университет. – Йошкар-Ола : Поволжский государственный технологический университет, 2023. – 272 с. : ил., табл., схем. – Режим доступа: по подписке. – URL: <https://biblioclub.ru/index.php?page=book&id=714624>.
3. Ясницкий, Л. Н. Интеллектуальные системы : учебник / Л. Н. Ясницкий. – 2-е изд. – Москва : Лаборатория знаний, 2020. – 224 с. : ил., табл., схем. – (Учебник для высшей школы). – Режим доступа: по подписке. – URL: <https://biblioclub.ru/index.php?page=book&id=712949>.

3.1.2. Дополнительная

1. Машинное обучение : учебник : [16+] / Е. Ю. Бутырский, В. В. Цехановский, Н. А. Жукова [и др.]. – Москва : Директ-Медиа, 2023. – 368 с. : ил., табл., схем., граф. – Режим доступа: по подписке. – URL: <https://biblioclub.ru/index.php?page=book&id=701807>.

3.2. Профессиональные базы данных и информационные справочные системы

Наименование ресурса	Адрес	Тип доступа
Электронно-библиотечные системы		
ЭБС «Университетская библиотека онлайн»	https://biblioclub.ru/	Индивидуальный неограниченный доступ после регистрации
КиберЛенинка, российская научная электронная библиотека	https://cyberleninka.ru/	Открытый ресурс
Профессиональные базы данных и информационные справочные системы		
eLIBRARY.RU, российский информационно-аналитический портал	https://elibrary.ru/	Открытый ресурс

3.3. Сетевые ресурсы

Наименование ресурса	Адрес

3.4. Методическое обеспечение дисциплины

3.4.1. Методические указания для преподавателя

Изучение дисциплины проводится в форме лекций, практических занятий, организации самостоятельной работы студентов, консультаций.

Лекции реализуются через изложение учебного материала с возможным мультимедийным сопровождением. Основными целями лекции являются системное освещение ключевых понятий и положений по соответствующей теме, обзор и оценка существующей проблематики, её методологических и социокультурных оснований, возможных вариантов решения, предложение методических рекомендаций для дальнейшего изучения курса, в том числе литературы и источников. Лектор должен стимулировать обучающихся к участию в обсуждении вопросов лекционного занятия, к высказыванию собственной точки зрения по обсуждаемой проблеме. Главное назначение лекции — обеспечить теоретическую основу обучения, развить интерес к учебной деятельности и конкретной учебной дисциплине, сформировать у обучающихся ориентиры для самостоятельной работы над курсом.

Практические занятия направлены на развитие самостоятельности обучающихся в исследовании изучаемых вопросов и приобретение умений и навыков. Практические занятия традиционно проводятся в форме обсуждения проблемных вопросов в группе при активном участии обучающихся. Они способствуют углубленному изучению наиболее фундаментальных и сложных проблем курса, служат важной формой анализа и синтеза исследуемого материала, а также подведения итогов самостоятельной работы обучающихся, стимулируя развитие профессиональной компетентности, навыков и умений. На практических занятиях обучающиеся учатся работать с научной литературой, чётко и доходчиво излагать проблемы и предлагать варианты их решения, аргументировать свою позицию, оценивать и критиковать позиции других, свободно публично высказывать свои мысли и суждения, грамотно вести полемику и представлять результаты собственных исследований. Практические занятия проводятся в форме устных и письменных опросов, диспута, тестирования, выполнения заданий, обсуждения докладов, выполнения контрольных заданий и пр.

Результаты работы на практических занятиях учитываются преподавателем при выставлении итоговой оценки по данной дисциплине. На усмотрение преподавателя обучающиеся, активно отвечающие на занятиях и выполняющие рекомендации преподавателя при подготовке к ним, могут получить повышающий балл к своей оценке в рамках промежуточной аттестации.

3.4.2. Методические рекомендации по самостоятельной работе обучающихся

Самостоятельная работа обучающихся выполняется по заданию и при методическом руководстве преподавателя, но без его непосредственного участия. Самостоятельная работа подразделяется на самостоятельную работу на аудиторных занятиях и на внеаудиторную самостоятельную работу. Самостоятельная работа обучающихся включает как полностью самостоятельное освоение отдельных тем (разделов) дисциплины, так и проработку тем (разделов), осваиваемых во время аудиторной работы. Во время самостоятельной работы обучающиеся читают и конспектируют учебную, научную и справочную литературу, выполняют задания, направленные на закрепление знаний и отработку умений и навыков, готовятся к текущему и промежуточному контролю по дисциплине.

Самостоятельная работа включается в общую трудоёмкость учебной нагрузки обучающегося.

Реализация поставленной цели предполагает решение следующих задач:

- освоение и расширение теоретических знаний по изучаемой дисциплине;
- систематизация и закрепление полученных теоретических знаний и практических навыков;
- формирование умений по поиску и использованию нормативной, справочной и специальной литературы, а также других источников информации;
- развитие познавательных способностей и активности, творческой инициативы, самостоятельности, ответственности и организованности;
- формирование самостоятельности мышления, способностей к саморазвитию, самообразованию, самосовершенствованию и самореализации;
- развитие научно-исследовательских навыков;
- формирование умения решать практические задачи (в профессиональной деятельности), используя приобретённые знания, способности и навыки.

Самостоятельная работа должна быть выполнена индивидуально или являться частью коллективной работы (в случае выполнения группового задания в работе делается соответствующая оговорка).

3.5. Материально-техническое обеспечение дисциплины

3.5.1. Аудитории для проведения занятий

Специальные помещения представляют собой учебные аудитории для проведения занятий лекционного типа, занятий семинарского типа, выполнения курсовых работ, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы и помещения для хранения и профилактического обслуживания учебного оборудования.

Специальные помещения укомплектованы учебной мебелью, в том числе мебелью для преподавателя дисциплины, учебной доской.

Обучающиеся из числа лиц с ограниченными возможностями здоровья обеспечиваются электронными образовательными ресурсами в формах, адаптированных к ограничениям их здоровья.

Материально-техническое оснащение учебных аудиторий конкретизировано на официальном сайте Университета в подразделе «Материально-техническое обеспечение и оснащённость образовательного процесса. Доступная среда» раздела «Сведения об образовательной организации».

3.5.2. Оборудование и технические средства обучения

Специальные помещения укомплектованы демонстрационным оборудованием (мультимедийный проектор, экран, компьютер, звуковые колонки).

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети Интернет и обеспечением доступа в электронную информационно-образовательную среду Университета.

Электронно-библиотечная система (электронная библиотека) и электронная информационно-образовательная среда обеспечивают одновременный доступ не менее 25 процентов обучающихся.

3.5.3. Программное обеспечение

Наименование	Сведения о лицензии
Moodle, среда дистанционного обучения	GNU GPL, свободно распространяемое с открытым исходным кодом

4. ОЦЕНОЧНЫЕ СРЕДСТВА

4.1. Методика проведения текущего контроля успеваемости и промежуточной аттестации

Оценивание уровня учебных достижений обучающихся по дисциплине осуществляется в виде текущего контроля успеваемости и промежуточной аттестации. Фонд оценочных средств по дисциплине включает оценочные материалы, направленные на проверку освоения компетенций, в том числе знаний, умений и навыков. Фонд оценочных средств включает оценочные средства текущего контроля и оценочные средства промежуточной аттестации.

4.1.1. Балльно-рейтинговая карта дисциплины

Виды работы	Виды контроля, критерии оценки и количество баллов
Текущий контроль	
Аудиторная работа (0-20 баллов)	Посещение занятий и участие в работе: 10 баллов — посещение не менее 50% занятий 15 баллов — посещение 60-80% занятий, участие в обсуждениях материала 20 баллов — посещение 90-100% занятий, активное участие в обсуждениях материала
Самостоятельная работа (0-40 баллов)	
Промежуточная аттестация	
Контрольное задание (0-40 баллов)	Выполнение письменного контрольного задания в СДО Moodle

4.1.2. Шкала перевода рейтинговых баллов в пятибалльную систему оценивания

Общее количество баллов	Оценка сформированности компетенций	Оценка результатов обучения по дисциплине	Оценка ECTS
0–34	Компетенции не сформированы.	неудовлетворительно (не зачтено)	F
Теоретическое содержание не освоено, практические навыки не сформированы, большинство учебных заданий не выполнено, качество их выполнения минимальное, все задания содержат грубые ошибки. Обучающийся не готов решать типовые профессиональные задачи.			
35–49	Уровень владения компетенциями недостаточный для их формирования в результате обучения по дисциплине.	неудовлетворительно (не зачтено)	FX
Теоретическое содержание освоено частично, практические навыки не сформированы, большинство учебных заданий не выполнено, качество их выполнения минимальное, большинство заданий выполнено с ошибками. Обучающийся не готов решать типовые профессиональные задачи.			
50–59	Уровень владения компетенциями	удовлетворительно	E

Общее количество баллов	Оценка сформированности компетенций	Оценка результатов обучения по дисциплине	Оценка ECTS
	посредственный для их формирования в результате обучения по дисциплине.	(зачтено)	
Теоретическое содержание освоено частично, практические навыки сформированы фрагментарно, многие учебные задания не выполнены, качество их выполнения минимальное, многие задания выполнены с ошибками. Обучающийся готов решать типовые профессиональные задачи.			
60–69	Уровень владения компетенциями удовлетворительный для их формирования в результате обучения по дисциплине.	удовлетворительно (зачтено)	D
Теоретическое содержание освоено частично, пробелы не носят существенного характера, практические навыки в основном сформированы, большинство учебных заданий выполнено, качество их выполнения удовлетворительное, некоторые задания выполнены с ошибками. Обучающийся готов решать типовые профессиональные задачи.			
70–89	Уровень владения компетенциями преимущественно высокий для их формирования в результате обучения по дисциплине.	хорошо (зачтено)	C
Теоретическое содержание освоено полностью, некоторые практические навыки сформированы недостаточно, все учебные задания выполнены, качество их выполнения высокое, некоторые задания выполнены с ошибками. Обучающийся готов решать типовые и ситуативные профессиональные задачи.			
90–94	Уровень владения компетенциями высокий для их формирования в результате обучения по дисциплине.	отлично (зачтено)	B
Теоретическое содержание освоено полностью, практические навыки сформированы, все учебные задания выполнены, качество их выполнения близко к максимальному, однако есть несколько незначительных ошибок. Обучающийся готов эффективно решать типовые и ситуативные профессиональные задачи, в том числе повышенного уровня сложности.			
95–100	Уровень владения компетенциями превосходный для их формирования в результате обучения по дисциплине.	отлично (зачтено)	A
Теоретическое содержание освоено полностью, практические навыки сформированы, все учебные задания выполнены, качество их выполнения близко к максимальному. Обучающийся готов эффективно решать типовые и ситуативные профессиональные задачи, в том числе повышенного уровня сложности, и способен разрабатывать новые решения.			

4.2. Оценочные средства текущего контроля

4.2.1. Вопросы для подготовки к семинарским занятиям

Раздел 1. Введение в интеллектуальный анализ данных и инженерию знаний

1. Опишите типовую архитектуру интеллектуальной системы в гуманитарной сфере. В чем ее принципиальное отличие от классических информационных систем?

2. Раскройте соотношение понятий в триаде «данные — информация — знания». Приведите примеры трансформации языковых данных в лингвистические знания.
3. Охарактеризуйте логическую и продукционную модели представления знаний. Каковы ограничения их применения для описания семантики естественного языка?
4. В чем заключается специфика фреймовой модели представления знаний? Как фреймы Марвина Минского соотносятся с понятием падежных грамматик в лингвистике?
5. Опишите структуру и принципы построения семантических сетей. Как на их основе реализуется поиск лингвистических закономерностей?
6. Что такое онтология и граф знаний в контексте Semantic Web? Приведите примеры известных лингвистических онтологий (например, WordNet).
7. Сформулируйте ключевые направления, прикладные задачи и специфику развития компьютерной лингвистики на современном этапе.
8. Опишите архитектуру и математические основы систем автоматического распознавания речи (ASR). С какими трудностями сталкиваются разработчики при обработке устного дискурса?
9. Каковы технологические принципы и алгоритмы синтеза речи (TTS)? Как современные генеративные нейросети решают проблему естественности интонации?
10. Каким образом инженерия знаний помогает формализовать контекстуальные и экстралингвистические факторы при проектировании диалоговых интерфейсов?
11. В чем заключается разница между символическим (экспертным) и коннекционистским (нейросетевым) подходами к моделированию языковой способности?
12. Каковы критерии оценки эффективности графов знаний при решении задач семантического поиска в больших массивах неструктурированного контента?

Раздел 2. Архитектура информационного поиска и предобработка текстовых данных

1. Сравните булеву, векторную и вероятностную модели информационного поиска. В чем заключаются их достоинства и ограничения при работе с полнотекстовыми базами данных?
2. Каковы принципы и алгоритмы индексации текстовых массивов (прямой и инвертированный индексы)?
3. Опишите полный цикл предобработки неструктурированных языковых данных (Data Preprocessing Pipeline). Почему качество предобработки критично для ИИ-моделей?
4. В чем заключается разница между стеммингом (например, алгоритм Портера) и лемматизацией? В каких NLP-задачах предпочтительнее использовать каждый из методов?
5. Раскройте содержание этапа токенизации текста. С какими краевыми случаями (пунктуация, дефисы, сокращения) сталкиваются современные токенизаторы?
6. Каковы принципы фильтрации стоп-слов и очистки текстового шума? Как избыточное удаление стоп-слов может исказить результаты прагматического анализа текста?
7. Что такое морфологический парсинг (POS-tagging) и какие подходы (правила, статистические модели, нейросети) используются для снятия омонимии?
8. Опишите задачи и методы синтаксического парсинга текста (грамматики зависимостей и грамматики составляющих).
9. Проведите сравнительный анализ прикладных Python-библиотек NLTK и SpaCy: производительность, архитектура, поддержка русского языка и готовые пайплайны.
10. Как устроены современные токенизаторы подсловных единиц (BPE, WordPiece, SentencePiece) и почему они вытеснили токенизацию по пробелам в моделях ИИ?
11. Каким образом особенности флективных языков (в частности, русского) усложняют этап лемматизации и синтаксического анализа по сравнению с аналитическими языками?
12. Что такое регулярные выражения (re в Python) и как они интегрируются в современные конвейеры предварительной обработки лингвистических данных?

Раздел 3. Методология и алгоритмы машинного обучения в лингвистике

1. Дайте определение машинному обучению (Machine Learning). В чем состоят фундаментальные различия между парадигмами обучения с учителем, без учителя и с подкреплением?
2. Охарактеризуйте метрику Accuracy (точность). В каких случаях (например, при сильном дисбалансе классов в корпусе текстов) эта метрика становится нерепрезентативной?
3. Раскройте математический и лингвистический смысл метрик Precision (точность поиска/классификации) и Recall (полнота). Как соблюсти баланс между ними?
4. Что такое F-мера (в частности, F1-score)? Каковы математические свойства среднего гармонического и почему для оценки NLP-систем используют именно его, а не среднее арифметическое?
5. Опишите принцип работы кривой ROC-AUC и матрицы ошибок (Confusion Matrix) при оценке качества бинарной классификации текстов.
6. Как устроен алгоритм Наивного Байесовского классификатора (Naive Bayes)? Почему, несмотря на «наивное» предположение о независимости признаков, он эффективен в задаче фильтрации спама?
7. Объясните принцип работы Метода опорных векторов (SVM) применительно к классификации текстовых документов высокой размерности.
8. Каковы математические основы Логистической регрессии? Как логит-трансформация позволяет предсказывать вероятность отнесения текста к определенному классу?
9. Что такое анализ тональности текста (Sentiment Analysis)? Какие лингвистические маркеры (ирония, отрицание, сарказм) вызывают наибольшие трудности у классических алгоритмов ML?
10. В чем разница между задачами классификации и кластеризации текстов? Какие алгоритмы обучения без учителя (например, K-Means, LDA) применяются для тематического моделирования?
11. Что такое переобучение (Overfitting) модели машинного обучения и какие методы регуляризации и валидации данных (кросс-валидация) используются для борьбы с ним?
12. Каким образом лингвист-аналитик осуществляет отбор и конструирование признаков (Feature Engineering) для классических алгоритмов машинного обучения?

Раздел 4. Векторные представления слов: статические и контекстуальные эмбединги

1. В чем заключается суть гипотезы дистрибутивной семантики (Дж. Фёрс)? Как фраза «узнай слово по его окружению» реализована в компьютерной лингвистике?
2. Опишите разреженные модели векторного представления слов (One-Hot Encoding, Bag of Words). Каковы их главные недостатки (размерность, отсутствие семантических связей)?
3. Как устроена модель взвешивания признаков TF-IDF? Опишите математический смысл компонентов Term Frequency и Inverse Document Frequency.
4. Опишите архитектуру статической модели эмбедингов Word2Vec. В чем разница между режимами обучения Continuous Bag-of-Words (CBOW) и Skip-gram?
5. Каковы особенности моделей FastText и GloVe? Как FastText решает проблему слов, отсутствующих в словаре (Out-of-Vocabulary), с помощью n-грамм символов?
6. Продемонстрируйте суть векторной семантики на классическом примере алгебры слов: «король» – «мужчина» + «женщина» = «королева». Как косинусное расстояние определяет семантическую близость?
7. Сформулируйте ключевые ограничения статических моделей эмбедингов. Как они справляются (или не справляются) с проблемой лексической многозначности и омонимии?
8. В чем принципиальное отличие контекстуальных эмбедингов (ELMo, BERT) от статических? Как меняется вектор слова «замок» в контекстах архитектуры и анатомии?
9. Опишите революционное значение архитектуры Transformer. Как устроен механизм внутреннего внимания (Self-Attention) и почему он позволил отказаться от рекуррентных сетей (RNN/LSTM)?
10. Объясните логику предобучения модели BERT (задачи Masked Language Modeling и Next Sentence Prediction). В чем заключается двунаправленность этой модели?

11. Что такое задача распознавания именованных сущностей (NER) и как предобученные трансформерные модели автоматизируют извлечение фактов из юридических или медицинских текстов?
12. Каким образом лингвист может дообучить (Fine-Tuning) готовую модель BERT под узкоспециализированную задачу своей проектной группы?

Раздел 5. Большие языковые модели (LLM) и генеративный искусственный интеллект

1. Опишите эволюцию больших языковых моделей (LLM) от архитектуры Vanilla Transformer до современных семейств моделей (GPT, LLaMA, YandexGPT, Claude).
2. Назовите ключевые возможности и технологические ограничения LLM. Чем обусловлено резкое возрастание качества генерации при увеличении числа параметров модели (Emergent Abilities)?
3. Что такое феномен «галлюцинаций» больших языковых моделей? Каковы лингвистические и архитектурные причины генерации фактологически неверного текста?
4. Раскройте понятие инженерии запросов (Prompt Engineering). Каковы основные принципы составления эффективных промптов для решения профессиональных задач лингвиста?
5. Сравните подходы к проектированию запросов: Zero-shot, Few-shot и Chain-of-Thought (цепочка рассуждений). В каких кейсах их применение наиболее оправданно?
6. Какие существуют методы автоматической валидации, верификации и фактчекинга результатов работы генеративных нейросетей (например, архитектура RAG — Retrieval-Augmented Generation)?
7. В чем суть этапа выравнивания моделей (Alignment) с помощью методов RLHF (обучение с подкреплением на основе отзывов людей) и RLAIIF (на основе отзывов ИИ)?
8. Опишите сценарии интеграции генеративных нейросетей в профессиональную деятельность цифрового лингвиста (автоматическое реферирование, разметка корпусов, генерация глоссариев).
9. С какими этическими, правовыми и концептуальными вызовами сталкивается академическое сообщество при использовании LLM для генерации научных и учебных текстов?
10. Что такое контекстное окно большой языковой модели и как его размер влияет на способность ИИ анализировать длинные документы и художественные произведения?
11. Каковы особенности применения LLM в качестве интеллектуальных агентов, способных вызывать внешние инструменты (API, калькуляторы, интерпретаторы кода Python)?
12. Опишите современные методы оценки (бенчмарки) качества работы LLM в различных лингвистических и логических доменах (например, MMLU, семейство тестов для русского языка).

Критерии оценки работы на практическом занятии

Критерии	Максимальное количество баллов за занятие
Устный опрос, коллоквиум	
Основные теоретические положения по вопросу раскрыты. Имеются элементы обоснования выводов. Имеются элементы систематизации информации, факты применения профессиональной терминологии. Очевидно использование источников рекомендованной литературы.	5 баллов

4.2.2. Темы докладов

1. Эволюция моделей представления знаний: от классических экспертных систем к современным графам знаний.
2. Онтологический инжиниринг в лингвистике: принципы проектирования предметных областей в редакторе Protégé.
3. Технологии распознавания эмоций в речевом сигнале (Speech Emotion Recognition) на базе интеллектуальных систем ASR.
4. Сравнительный анализ векторной и вероятностной моделей информационного поиска в современных поисковых архитектурах.
5. Проектирование оптимального конвейера предобработки (Data Preprocessing Pipeline) для зашумленного текстового контента социальных сетей.
6. Проблема снятия грамматической и лексической омонимии в автоматизированных морфологических парсерах.
7. Применение классических алгоритмов машинного обучения (SVM, Naïve Bayes) для автоматического определения жанра текста.
8. Анализ тональности (Sentiment Analysis) и извлечение аспектов (Aspect-Based Sentiment Analysis) в отзывах пользователей: лингвистический подход.
9. Математические и концептуальные основы дистрибутивной семантики: от частотных матриц совместной встречаемости к Word2Vec.
10. Преодоление проблемы Out-of-Vocabulary (OOV) слов в компьютерной лингвистике: технологии FastText и Byte Pair Encoding (BPE).
11. Архитектура Transformer и механизм внимания (Attention): технологический прорыв в области обработки естественного языка.
12. Модель BERT и ее модификации (RoBERTa, ALBERT, SciBERT): опыт применения в прикладных лингвистических исследованиях.
13. Извлечение именованных сущностей (NER) и связывание сущностей (Entity Linking) в задачах автоматического наполнения баз знаний.
14. Феномен эмерджентных свойств (Emergent Abilities) больших языковых моделей при масштабировании параметров.
15. Технологические подходы к снижению уровня «галлюцинаций» больших языковых моделей.
16. Методология Prompt Engineering: продвинутые техники управления генерацией контента (Chain-of-Thought, Tree of Thoughts).
17. Использование архитектуры RAG (Retrieval-Augmented Generation) для построения экспертных систем на закрытых лингвистических корпусах.
18. Трансформация профессии лингвиста в эпоху доминирования генеративного искусственного интеллекта: вызовы и перспективы.

Шкала и критерии оценки доклада

Критерии	Показатели	Баллы
1. Степень раскрытия сущности проблемы	– соответствие теме доклада; – полнота и глубина раскрытия основных понятий; – умение работать с литературой, систематизировать и структурировать материал; – умение обобщать, сопоставлять различные точки зрения по рассматриваемому вопросу, аргументировать основные положения и выводы.	70

Критерии	Показатели	Баллы
2. Обоснованность выбора источников	– круг, полнота использования литературных источников по теме; – привлечение новейших работ (журнальные публикации, материалы сборников научных трудов и т.д.).	15
3. Изложение	– литературный стиль.	15

Доклад оценивается по 100 балльной шкале, баллы переводятся в оценки успеваемости следующим образом:

90 – 100 баллов – «отлично»;

70 – 89 баллов – «хорошо»;

50 – 69 баллов – «удовлетворительно»;

менее 50 баллов – «неудовлетворительно».

4.3. Оценочные средства промежуточной аттестации

4.3.1. Контрольные задания

4.3.2. Ключи к контрольным заданиям

Шкала и критерии оценки текущего тестирования

Число правильных ответов	Оценка
90-100% правильных ответов	Оценка «отлично»
70-89% правильных ответов	Оценка «хорошо»
50-69% правильных ответов	Оценка «удовлетворительно»
Менее 50% правильных ответов	Оценка «неудовлетворительно»

5. ОСОБЕННОСТИ РЕАЛИЗАЦИИ ДИСЦИПЛИНЫ В ОТНОШЕНИИ ЛИЦ ИЗ ЧИСЛА ИНВАЛИДОВ И ЛИЦ С ОГРАНИЧЕННЫМИ ВОЗМОЖНОСТЯМИ ЗДОРОВЬЯ

Профессорско-преподавательский состав знакомится с психолого-физиологическими особенностями обучающихся инвалидов и лиц с ограниченными возможностями здоровья, индивидуальными программами реабилитации инвалидов (при наличии). При необходимости осуществляется дополнительная поддержка преподавания психологами, социальными работниками, прошедшими подготовку ассистентами.

В соответствии с методическими рекомендациями Министерства науки и высшего образования Российской Федерации, в курсе предполагается использовать социально-активные и рефлексивные методы обучения, технологии социокультурной реабилитации с целью оказания помощи в установлении полноценных межличностных отношений с другими студентами, создания комфортного психологического климата в студенческой группе. Подбор и разработка учебных материалов производятся с учётом предоставления материала в различных формах: аудиальной, визуальной, с использованием специальных технических средств и информационных систем.

Согласно требованиям к порядку реализации образовательной деятельности в отношении инвалидов и лиц с ОВЗ, установленным Министерством науки и высшего образования Российской Федерации, необходимо иметь в виду, что:

1) инвалиды и лица с ОВЗ по зрению имеют право присутствовать на занятиях вместе с ассистентом, оказывающим обучающемуся необходимую помощь;

2) инвалиды и лица с ОВЗ по слуху имеют право на использование звукоусиливающей аппаратуры.

При проведении промежуточной аттестации по дисциплине обеспечивается соблюдение следующих общих требований:

- проведение аттестации для инвалидов в одной аудитории совместно с обучающимися, не являющимися инвалидами, если это не создает трудностей для инвалидов и иных обучающихся при промежуточной аттестации;

- присутствие в аудитории ассистента (ассистентов), оказывающего обучающимся инвалидам необходимую техническую помощь с учётом их индивидуальных особенностей (занять рабочее место, передвигаться, прочитать и оформить задание, общаться с экзаменатором);

- пользование необходимыми обучающимся инвалидам техническими средствами при прохождении промежуточной аттестации с учётом их индивидуальных особенностей;

- обеспечение возможности беспрепятственного доступа обучающихся инвалидов в аудитории, туалетные и другие помещения, а также их пребывания в указанных помещениях.

По письменному заявлению обучающегося инвалида продолжительность прохождения испытания промежуточной аттестации (зачёта, экзамена, и др.) обучающимся инвалидом может быть увеличена по отношению к установленной продолжительности его сдачи:

- продолжительность сдачи испытания, проводимого в письменной форме, — не более чем на 90 минут;

- продолжительность подготовки обучающегося к ответу, проводимому в устной форме, — не более чем на 20 минут.

В зависимости от индивидуальных особенностей обучающихся с ОВЗ Университет обеспечивает выполнение следующих требований при проведении аттестации:

а) для слепых:

- задания и иные материалы для прохождения промежуточной аттестации оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением для слепых, либо зачитываются ассистентом;

- письменные задания выполняются обучающимися на бумаге рельефно-точечным шрифтом Брайля или на компьютере со специализированным программным обеспечением для слепых, либо надиктовываются ассистенту;

- при необходимости обучающимся предоставляется комплект письменных принадлежностей и бумага для письма рельефно-точечным шрифтом Брайля, компьютер со специализированным программным обеспечением для слепых;

б) для слабовидящих:

- задания и иные материалы для сдачи экзамена оформляются увеличенным шрифтом;

- обеспечивается индивидуальное равномерное освещение не менее 300 люкс;

- при необходимости обучающимся предоставляется увеличивающее устройство, допускается использование увеличивающих устройств, имеющихся у обучающихся;

в) для глухих и слабослышащих, с тяжелыми нарушениями речи:

- обеспечивается наличие звукоусиливающей аппаратуры коллективного пользования, при необходимости обучающимся предоставляется звукоусиливающая аппаратура индивидуального пользования;

- по их желанию испытания проводятся в письменной форме;

г) для лиц с нарушениями опорно-двигательного аппарата (тяжелыми нарушениями двигательных функций верхних конечностей или отсутствием верхних конечностей):

- письменные задания выполняются обучающимися на компьютере со специализированным программным обеспечением или надиктовываются ассистенту;

- по их желанию испытания проводятся в устной форме.

О необходимости обеспечения специальных условий для проведения аттестации обучающийся должен сообщить письменно не позднее, чем за 10 дней до начала аттестации. К заявлению прилагаются документы, подтверждающие наличие у обучающегося индивидуальных особенностей (при отсутствии указанных документов в организации).

При необходимости для обучающихся с инвалидностью процедура оценивания результатов обучения может проводиться в несколько этапов.