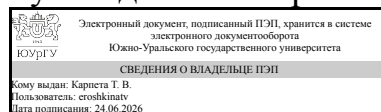


ЮЖНО-УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УТВЕРЖДАЮ:
Руководитель направления



Т. В. Карпета

РАБОЧАЯ ПРОГРАММА

дисциплины ФД.04 Основы построения RAG систем
для направления 02.04.02 Фундаментальная информатика и информационные технологии

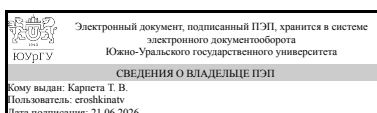
уровень Магистратура

форма обучения очная

кафедра-разработчик Прикладная математика и программирование

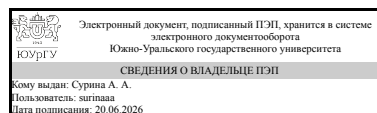
Рабочая программа составлена в соответствии с ФГОС ВО по направлению подготовки 02.04.02 Фундаментальная информатика и информационные технологии, утверждённым приказом Минобрнауки от 23.08.2017 № 811

Зав.кафедрой разработчика,
к.физ.-мат.н.



Т. В. Карпета

Разработчик программы,
старший преподаватель



А. А. Сурина

1. Цели и задачи дисциплины

Целью освоения дисциплины является формирование у магистрантов глубоких теоретических знаний и устойчивых практических навыков для исследования и разработки масштабируемых архитектур интеллектуальных информационных систем, базирующихся на технологии генерации, дополненной поиском (Retrieval-Augmented Generation, RAG). Для достижения данной цели в рамках курса решается комплекс задач: фундаментальное изучение математического аппарата векторного представления семантики данных и алгоритмов индексирования; освоение методов проектирования конвейеров поиска знаний с применением высокопроизводительных векторных баз данных; получение инженерного опыта интеграции как облачных, так и локальных больших языковых моделей (LLM) с внешними источниками данных; а также овладение строгими аналитическими методологиями количественной оценки, валидации и оптимизации качества генерируемых ответов.

Краткое содержание дисциплины

Содержание дисциплины охватывает полный жизненный цикл проектирования, разработки и тестирования современных RAG-архитектур. Учебный курс включает детальное изучение базовых и продвинутых этапов построения RAG, методов семантической векторизации текста, а также принципов функционирования векторных хранилищ (в частности, FAISS и Qdrant) и их обслуживание с фокусом на алгоритмы приближенного поиска ближайших соседей (HNSW, IVF). Особое внимание в программе уделено усложненным архитектурным паттернам: реализации гибридного поиска (интеграция разреженных матриц BM25 и плотных векторных представлений), механизмам кросс-энтропийного переранжирования (re-ranking) и алгоритмам трансформации пользовательского ввода (query rewriting, multi-query). На практике обучающиеся осваивают разработку пайплайнов с применением ведущих индустриальных фреймворков (LangChain, LlamaIndex), внедрение локальных инференс-движков (Ollama, vLLM) для обеспечения конфиденциальности обработки данных, методы оптимизации производительности (top-k поиск, кэширование) и комплексную оценку качества интеллектуальных систем с использованием специализированных метрик фреймворка RAGAS.

2. Компетенции обучающегося, формируемые в результате освоения дисциплины

Планируемые результаты освоения ОП ВО (компетенции)	Планируемые результаты обучения по дисциплине
ПК-2 Способен исследовать и разрабатывать архитектуры интеллектуальных информационных систем для различных предметных областей на основе комплексов методов и инструментальных средств систем искусственного интеллекта	Знает: этапы RAG, embedding, метрики схожести, векторные БД (FAISS, Qdrant) с HNSW/IVF, гибридный поиск (BM25+вектор), re-ranking, паттерны RAG, фреймворки (LangChain, LlamaIndex), метрики RAGAS. Умеет: выбирать архитектуру RAG под предметную область, проектировать чанкинг и embedding-модель, реализовывать ретривер с гибридным поиском и переранжированием, составлять промпты с контекстом для LLM и оценивать качество системы по метрикам полноты и точности.

	Имеет практический опыт: сборки RAG-пайплайна на Python с использованием LangChain или LlamaIndex, интеграции локальных (Ollama, vLLM) и облачных LLM, оптимизации поиска (top-k, порог схожести, кэширование эмбеддингов), отладки и улучшения RAG (query rewriting, multi-query, анализ нерелевантного контекста), а также документирования архитектуры и формирования отчёта по оценке качества RAG-системы.
--	---

3. Место дисциплины в структуре ОП ВО

Перечень предшествующих дисциплин, видов работ учебного плана	Перечень последующих дисциплин, видов работ
Нет	Не предусмотрены

Требования к «входным» знаниям, умениям, навыкам студента, необходимым при освоении данной дисциплины и приобретенным в результате освоения предшествующих дисциплин:

Нет

4. Объём и виды учебной работы

Общая трудоемкость дисциплины составляет 3 з.е., 108 ч., 38,25 ч. контактной работы

Вид учебной работы	Всего часов	Распределение по семестрам в часах
		Номер семестра
		3
Общая трудоёмкость дисциплины	108	108
<i>Аудиторные занятия:</i>	32	32
Лекции (Л)	16	16
Практические занятия, семинары и (или) другие виды аудиторных занятий (ПЗ)	16	16
Лабораторные работы (ЛР)	0	0
<i>Самостоятельная работа (СРС)</i>	69,75	69,75
Подготовка к практическим занятиям, программная реализация пайплайнов RAG, развертывание векторных хранилищ и локальных инференс-серверов, расчет метрик качества и оформление технических отчетов.	34,75	34.75
Изучение основной и дополнительной литературы, анализ научных статей по архитектурам RAG, дистрибутивной семантике и алгоритмам масштабируемого векторного поиска.	25	25
Подготовка к промежуточной аттестации (зачету)	10	10
Консультации и промежуточная аттестация	6,25	6,25
Вид контроля (зачет, диф.зачет, экзамен)	-	зачет

5. Содержание дисциплины

№ раздела	Наименование разделов дисциплины	Объем аудиторных занятий по видам в часах			
		Всего	Л	ПЗ	ЛР
1	Теоретико-математические основы архитектуры RAG и семантической векторизации	8	4	4	0
2	Проектирование векторных хранилищ и алгоритмы масштабируемого поиска	8	4	4	0
3	Продвинутые паттерны поиска и интеграция больших языковых моделей (LLM)	8	4	4	0
4	Оптимизация, развертывание и методология оценки качества RAG-систем	8	4	4	0

5.1. Лекции

№ лекции	№ раздела	Наименование или краткое содержание лекционного занятия	Кол- во часов
1	1	Архитектура и эволюция систем RAG. Концептуальная схема пайплайна RAG. Проблема галлюцинаций в LLM и методы её решения через добавление внешнего контекста. Жизненный цикл обработки запроса: индексация, извлечение, генерация.	2
2	1	Математический аппарат семантической векторизации. Дистрибутивная семантика и модели погружения. Архитектуры моделей кодировщиков. Метрики векторной близости: косинусное сходство, L2-расстояние, скалярное произведение.	2
3	2	Векторные базы данных. Архитектура и назначение специализированных СУБД для векторов. Сравнительный анализ решений: in-memory библиотеки (FAISS) и распределенные хранилища (Qdrant). Структуры данных для хранения.	2
4	2	Алгоритмы приближенного поиска ближайших соседей (ANN). Проблемы размерности в задачах KNN. Инвертированный индекс (IVF) и методы квантования. Иерархический навигационный граф малого мира (HNSW): параметры графа и их влияние на recall/latency.	2
5	3	Индустриальные фреймворки и интеграция LLM. Архитектурные паттерны оркестрации в LangChain и LlamaIndex. Различия подходов к абстракции данных. Подключение облачных и локальных (Ollama, vLLM) больших языковых моделей.	2
6	3	Гибридный поиск и кросс-энтропийное переранжирование. Ограничения плотного векторного поиска. Разреженный поиск (BM25). Ансамблирование результатов (Hybrid Search, RRF). Модели Re-ranking (Cross-Encoders) для уточнения релевантности.	2
7	4	Трансформация запросов и оптимизация контекста. Стратегии предварительной обработки ввода: Query Rewriting, Multi-query (генерация суб-запросов), Step-back prompting. Методы работы с контекстным окном LLM.	2
8	4	Метрики качества и производительности RAG. Системные оптимизации: настройка top-k выборки, семантическое кэширование эмбедингов. Фреймворк RAGAS: изолированная оценка этапов Retrieval и Generation.	2

5.2. Практические занятия, семинары

№ занятия	№ раздела	Наименование или краткое содержание практического занятия, семинара	Кол-во часов
1	1	Построение базового пайплайна токенизации и векторизации. Стратегии чанкинга текста (Recursive, Semantic chunking). Генерация эмбеддингов с использованием открытых моделей семейства HuggingFace.	2
2	1	Вычисление метрик векторного сходства. Реализация точного KNN поиска (Exact Search). Оценка вычислительной сложности и задержек поиска при программном росте объема корпуса документов.	2
3	2	Инициализация и настройка индексов в FAISS и Qdrant. Развертывание локальной инстанции Qdrant. Создание коллекций, настройка схем хранения, индексация чанков с метаданными.	2
4	2	Профилирование алгоритмов HNSW и IVF. Оценка компромисса между скоростью (QPS) и точностью при изменении гиперпараметров.	2
5	3	Оркестрация с помощью LangChain/LlamaIndex и локальный инференс. Настройка пайплайна извлечения. Развертывание локальной модели через Ollama/vLLM и её подключение к цепочке генерации ответов.	2
6	3	Настройка гибридного поиска и модуля Re-ranking. Реализация разреженного поиска алгоритмом BM25. Настройка весов объединения (BM25 + Dense). Интеграция легковесной модели переранжирования (BGE-Reranker).	2
7	4	Оптимизация пайплайна: Multi-query и кэширование. Программная реализация алгоритма Multi-query для расширения контекста поиска. Интеграция семантического кэша для снижения задержек повторяющихся запросов.	2
8	4	Комплексная оценка системы через фреймворк RAGAS. Формирование ground-truth датасета. Расчет метрик Faithfulness, Answer Relevance и Context Precision. Составление аналитического отчета по архитектуре.	2

5.3. Лабораторные работы

Не предусмотрены

5.4. Самостоятельная работа студента

Выполнение СРС			
Подвид СРС	Список литературы (с указанием разделов, глав, страниц) / ссылка на ресурс	Семестр	Кол-во часов
Подготовка к практическим занятиям, программная реализация пайплайнов RAG, развертывание векторных хранилищ и локальных инференс-серверов, расчет метрик качества и оформление технических отчетов.	Методические указания для СРС. Профессиональные базы данных (Qdrant, Milvus) и ресурсы (Ollama, vLLM, Hugging Face, DeepEval).	3	34,75
Изучение основной и дополнительной литературы, анализ научных статей по архитектурам RAG, дистрибутивной семантике и алгоритмам масштабируемого векторного поиска.	Основная литература 1–3. Дополнительная литература 1–3.	3	25
Подготовка к промежуточной аттестации (зачету)	Основная литература 1–3. Дополнительная литература 1–3.	3	10

6. Фонд оценочных средств для проведения текущего контроля успеваемости, промежуточной аттестации

Контроль качества освоения образовательной программы осуществляется в соответствии с Положением о балльно-рейтинговой системе оценивания результатов учебной деятельности обучающихся.

6.1. Контрольные мероприятия (КМ)

№ КМ	Се-местр	Вид контроля	Название контрольного мероприятия	Вес	Макс. балл	Порядок начисления баллов	Учитывается в ПА
1	3	Текущий контроль	КП-1	1	4	Баллы начисляются за соответствие работы критериям. Соответствие одному критерию 1 балл, максимум 4 балла за соответствие всем критериям: • корректная реализация стратегий чанкинга и генерации плотных векторных представлений (embeddings); • успешное развертывание и настройка коллекций в векторных хранилищах (FAISS или Qdrant); • проведение профилирования индексов (HNSW, IVF) с анализом метрик производительности (Recall@K, QPS); • уверенные ответы на теоретические вопросы при защите работы.	зачет
2	3	Текущий контроль	ПК-2	1	4	Баллы начисляются за соответствие работы критериям. Соответствие одному критерию 1 балл, максимум 4 балла за соответствие всем критериям: • безошибочная программная реализация гибридного поиска (разреженный алгоритм BM25 + плотный векторный поиск); • успешная интеграция кросс-энтропийной модели переранжирования (Re-ranking); • корректная оркестрация пайплайна средствами LangChain/LlamaIndex с подключением локальных инференс-движков (Ollama, vLLM); • аргументированное обоснование архитектурных решений в ходе защиты.	зачет
3	3	Текущий контроль	Баллы начисляются за соответствие работы критериям. Соответствие одному критерию 1 балл, максимум 4 балла за соответствие всем критериям:	1	4	Баллы начисляются за соответствие работы критериям. Соответствие одному критерию 1 балл, максимум 4 балла за соответствие всем критериям: • программная имплементация продвинутых методов трансформации запросов (Multi-query, Query Rewriting) и семантического кэширования; • оценка качества этапа извлечения	зачет

						(Retrieval Evaluation): расчет метрик Context Precision и Context Recall, анализ распределения оценок и влияние параметров векторной БД; • оценка качества этапа генерации (Generation Evaluation): расчет метрик Answer Faithfulness и Answer Relevance, сравнительный анализ влияния методов трансформации запросов на итоговое качество ответа; • строгое академическое оформление отчета и глубокий анализ полученных метрик (связь с архитектурой пайплайна).	
4	3	Промежуточная аттестация	Зачет	-	10	Баллы начисляются по результатам устного ответа на билет: • 5 баллов: исчерпывающий ответ на теоретический вопрос билета с использованием корректной терминологии машинного обучения (знание архитектуры, математического аппарата векторизации и алгоритмов ANN); • 5 баллов: успешное решение практико-ориентированной задачи (проектирование архитектуры RAG под заданную предметную область, выбор метрик и методов оптимизации).	зачет

6.2. Процедура проведения, критерии оценивания

Вид промежуточной аттестации	Процедура проведения	Критерии оценивания
зачет	На зачете происходит оценивание учебной деятельности обучающихся по дисциплине на основе полученных оценок за контрольно-рейтинговые мероприятия текущего контроля. Контрольное мероприятие промежуточной аттестации проводится во время зачета в виде устного опроса, если студенту не хватило баллов для выставления зачета по текущему контролю. В этом случае, при условии выполнения всех лабораторных работ, студенту выдается билет с двумя вопросами из разных тем курса, позволяющих оценить сформированность компетенций. Время на подготовку ответов не предусмотрено. Использование вспомогательных материалов при подготовке ответа не допускается. Ответ на каждый вопрос оценивается по 5 бальной шкале. Итоговой оценкой является сумма баллов за два вопроса. Если студент согласен с оценкой, полученной по результатам текущего контроля, то он может в день, предшествующий промежуточной аттестации дать свое согласие в личном кабинете. В случае явки студента на промежуточную аттестацию, давшего свое согласие на автомат в личном	В соответствии с пп. 2.5, 2.6 Положения

	кабинете, студент имеет право пройти мероприятия текущего контроля по дисциплине на промежуточной аттестации для улучшения своего рейтинга в день ее проведения. Снижение оценки в этом случае запрещено. Если студент не дал согласия в личном кабинете, то он может согласиться с оценкой лично на промежуточной аттестации в день ее проведения. Если студент не согласен с оценкой, то он имеет право пройти контрольно-рейтинговые мероприятия на промежуточной аттестации для улучшения своего рейтинга в день ее проведения. Фиксация результатов учебной деятельности по дисциплине проводится в день промежуточной аттестации на основе согласия студента, данного им в личном кабинете. При отсутствии согласия в журнале дисциплины фиксация результатов происходит при личном присутствии студента. Если студент не дал согласие в личном кабинете и не явился на промежуточную аттестацию – ему выставляется «неявка».	
--	--	--

6.3. Паспорт фонда оценочных средств

Компетенции	Результаты обучения	№ КМ			
		1	2	3	4
ПК-2	Знает: этапы RAG, embedding, метрики схожести, векторные БД (FAISS, Qdrant) с HNSW/IVF, гибридный поиск (BM25+вектор), re-ranking, паттерны RAG, фреймворки (LangChain, LlamaIndex), метрики RAGAS.	+	+		+
ПК-2	Умеет: выбирать архитектуру RAG под предметную область, проектировать чанкинг и embedding-модель, реализовывать ретривер с гибридным поиском и переранжированием, составлять промпты с контекстом для LLM и оценивать качество системы по метрикам полноты и точности.	+	+	+	+
ПК-2	Имеет практический опыт: сборки RAG-пайплайна на Python с использованием LangChain или LlamaIndex, интеграции локальных (Ollama, vLLM) и облачных LLM, оптимизации поиска (top-k, порог схожести, кэширование эмбеддингов), отладки и улучшения RAG (query rewriting, multi-query, анализ нерелевантного контекста), а также документирования архитектуры и формирования отчёта по оценке качества RAG-системы.		+	+	+

Типовые контрольные задания по каждому мероприятию находятся в приложениях.

7. Учебно-методическое и информационное обеспечение дисциплины

Печатная учебно-методическая документация

а) основная литература:

Не предусмотрена

б) дополнительная литература:

Не предусмотрена

в) отечественные и зарубежные журналы по дисциплине, имеющиеся в библиотеке:

Не предусмотрены

г) методические указания для студентов по освоению дисциплины:

1. Рекомендации по организации самостоятельной работы студентов

из них: учебно-методическое обеспечение самостоятельной работы студента:

1. Рекомендации по организации самостоятельной работы студентов

Электронная учебно-методическая документация

№	Вид литературы	Наименование ресурса в электронной форме	Библиографическое описание
1	Основная литература	ЭБС издательства Лань	Душкин, Р. В. RAG-системы. От теории к практике / Р. В. Душкин. — Москва : ДМК Пресс, 2026. — 286 с. — ISBN 978-5-93700-429-1. — Текст : электронный // Лань : электронно-библиотечная система. https://e.lanbook.com/book/521215
2	Основная литература	ЭБС издательства Лань	Пшехотская, Е. А. Информационный поиск : учебное пособие / Е. А. Пшехотская, Ю. Н. Филиппович. — Москва : Московский Политех, 2023. — 108 с. — ISBN 978-5-2760-2830-9. — Текст : электронный // Лань : электронно-библиотечная система. https://e.lanbook.com/book/482786
3	Основная литература	ЭБС издательства Лань	Грейнджер, Т. Поиск на основе искусственного интеллекта : руководство / Т. Грейнджер, Д. Тернбулл, М. Ирвин ; пер. с англ. И. Л. Люско. — Москва : ДМК Пресс, 2025. — 586 с. — ISBN 978-5-93700-180-1. — Текст : электронный // Лань : электронно-библиотечная система. https://e.lanbook.com/book/514904
4	Дополнительная литература	ЭБС издательства Лань	Гольдберг, Й. Нейросетевые методы в обработке естественного языка : руководство / Й. Гольдберг ; перевод с английского А. А. Слинкина. — Москва : ДМК Пресс, 2019. — 282 с. — ISBN 978-5-97060-754-1. — Текст : электронный // Лань : электронно-библиотечная система. https://e.lanbook.com/book/131704
5	Дополнительная литература	ЭБС издательства Лань	Халабия, Р. Ф. Математическое и программное обеспечение информационно-поисковых систем : учебное пособие / Р. Ф. Халабия, М. Л. Халабия, Л. В. Бунина. — Москва : РТУ МИРЭА, 2020. — 132 с. — Текст : электронный // Лань : электронно-библиотечная система. https://e.lanbook.com/book/167568
6	Дополнительная литература	ЭБС издательства Лань	Душкин, Р. В. Генеративный искусственный интеллект : руководство / Р. В. Душкин. — Москва : ДМК Пресс, 2025. — 228 с. — ISBN 978-5-93700-374-4. — Текст : электронный // Лань : электронно-библиотечная система. https://e.lanbook.com/book/514902

Перечень используемого программного обеспечения:

Нет

Перечень используемых профессиональных баз данных и информационных справочных систем:

Нет

8. Материально-техническое обеспечение дисциплины

Вид занятий	№ ауд.	Основное оборудование, стенды, макеты, компьютерная техника, предустановленное программное обеспечение, используемое для различных видов занятий
Лекции	333 (36)	Компьютер+проектор, персональные компьютеры
Практические занятия и семинары	333 (36)	Компьютер + проектор, персональные компьютеры