МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
"Казанский (Приволжский) федеральный университет"
Институт математики и механики им. Н.И. Лобачевского

УТВЕРЖДАЮ
Проректор по образовательной деятельности КФУ
_____ Е.А. Турилова
28 февраля 2025 г.

подписано электронно-цифровой подписью

Программа дисциплины

Введение в текстовую аналитику

Направление подготовки: 01.04.01 - Математика
Профиль подготовки: Анализ на многообразиях
Квалификация выпускника: магистр
Форма обучения: очное
Язык обучения: русский
Год начала обучения по образовательной программе: 2025

Содержание

Программу дисциплины разработал(а)(и): директор института математики и механики им. Н.И. Лобачевского Насрутдинов М.Ф. (директорат ИМиМ, Институт математики и механики им. Н.И. Лобачевского), Marat.Nasrutdinov@kpfu.ru

1. Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения ОПОП ВО

Обучающийся, освоивший дисциплину (модуль), должен обладать следующими компетенциями:

Шифр компетенции	Расшифровка приобретаемой компетенции
ПК-2	Способен активно участвовать в исследовании новых математических моделей в естественных науках

Обучающийся, освоивший дисциплину (модуль):

Должен знать:

- Основные задачи и методы текстовой аналитики.
- Принципы векторизации текста (TF-IDF, Word2Vec, Embeddings).
- Архитектуру и принципы работы трансформеров.
- Современные LLM (например, BERT, GPT, T5) и их особенности.

Должен уметь:

- Предобрабатывать текстовые данные для анализа.
- Применять библиотеки (NLTK, Transformers) для решения NLP-задач.
- Оценивать качество работы NLP-моделей.

Должен владеть:

- Навыками работы с Python-библиотеками для NLP.
- Навыками анализа и визуализации результатов текстовой аналитики

2. Место дисциплины (модуля) в структуре ОПОП ВО

Данная дисциплина (модуль) включена в раздел "Б1.В.ДВ.04.02 Дисциплины (модули)" основной профессиональной образовательной программы 01.04.01 "Математика (Анализ на многообразиях)" и относится к дисциплинам по выбору части ОПОП ВО, формируемой участниками образовательных отношений.

Осваивается на 1 курсе в 2 семестре.

3. Объем дисциплины (модуля) в зачетных единицах с указанием количества часов, выделенных на контактную работу обучающихся с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающихся

Общая трудоемкость дисциплины составляет 3 зачетных(ые) единиц(ы) на 108 часа(ов).

Контактная работа - 36 часа(ов), в том числе лекции - 12 часа(ов), практические занятия - 24 часа(ов), лабораторные работы - 0 часа(ов), контроль самостоятельной работы - 0 часа(ов).

Самостоятельная работа - 27 часа(ов).

Контроль (зачёт / экзамен) - 45 часа(ов).

Форма промежуточного контроля дисциплины: экзамен во 2 семестре.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1 Структура и тематический план контактной и самостоятельной работы по дисциплине (модулю)

N	Разделы дисциплины / модуля	Семестр	Виды и часы контактной работы, их трудоемкость (в часах)						Самостоятельная работа
			Лекции, всего	Лекции в эл. форме	Практические занятия, всего	Практические в эл. форме	Лабораторные работы, всего	Лабораторные в эл. форме	
N	Разделы дисциплины / модуля	Семестр	Виды и часы контактной работы, их трудоемкость (в часах)						Самостоятельная работа
			Лекции, всего	Лекции в эл. форме	Практические занятия, всего	Практические в эл. форме	Лабораторные работы, всего	Лабораторные в эл. форме	
1.	Тема 1. Введение в текстовую аналитику и NLP	2	2	0	4	0	0	0	5
2.	Тема 2. Предобработка текста и векторное представление	2	2	0	6	0	0	0	6
3.	Тема 3. Архитектура трансформеров и введение в LLM	2	4	0	8	0	0	0	8
4.	Тема 4. Применение LLM и промпт-инжиниринг	2	4	0	6	0	0	0	8
	Итого		12	0	24	0	0	0	27

4.2 Содержание дисциплины (модуля)

Тема 1. Введение в текстовую аналитику и NLP

Основные задачи обработки естественного языка (NLP): классификация, кластеризация, машинный перевод, генерация текстов. Жизненный цикл NLP-проекта.

Практика: Знакомство с окружением (Python, Jupyter Notebook). Базовые операции с текстом. Введение в библиотеки NLTK/spaCy. Использование Google Colab.

Тема 2. Предобработка текста и векторное представление

Токенизация, лемматизация, стемминг, очистка от стоп-слов. Существующие библиотеки для русского языка. Векторные представления: модель Мешок слов (Bag-of-Words), метрика TF-IDF, эмбеддинг слов (Word2Vec, GloVe).
Практика: Реализация полного пайплайна предобработки текста. Создание и визуализация эмбеддингов.

Тема 3. Архитектура трансформеров и введение в LLM

Принцип внимания (Attention Mechanism). Архитектура Encoder-Decoder. Модель Transformer: ключевые компоненты (Self-Attention, Feed-Forward Networks). Обзор семейств моделей: BERT (encoder), GPT (decoder), T5 (encoder-decoder).
Практика: Знакомство с токенизацией для современных моделей. Обзор соверменных моделей LLM.

Тема 4. Применение LLM и промпт-инжиниринг

Дообучение (Fine-tuning) vs. Промптинг. Принципы эффективного промпт-инжиниринга. Решение прикладных задач: чат-боты, анализ тональности, генерация контента. Инструменты: OpenAI API, Yandex GPT. Этические аспекты и безопасность LLM.

Практика: Написание промптов для различных задач. Создание простого чат-бота.

5. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине (модулю)

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.

Организация самостоятельной работы обучающихся регламентируется нормативными документами, учебно-методической литературой и электронными образовательными ресурсами, включая:

Порядок организации и осуществления образовательной деятельности по образовательным программам высшего образования - программам бакалавриата, программам специалитета, программам магистратуры (утвержден приказом Министерства науки и высшего образования Российской Федерации от 6 апреля 2021 года №245)

Письмо Министерства образования Российской Федерации №14-55-996ин/15 от 27 ноября 2002 г. "Об активизации самостоятельной работы студентов высших учебных заведений"

Устав федерального государственного автономного образовательного учреждения "Казанский (Приволжский) федеральный университет"

Правила внутреннего распорядка федерального государственного автономного образовательного учреждения высшего профессионального образования "Казанский (Приволжский) федеральный университет"

Локальные нормативные акты Казанского (Приволжского) федерального университета

6. Фонд оценочных средств по дисциплине (модулю)

Фонд оценочных средств по дисциплине (модулю) 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.

В фонде оценочных средств содержится следующая информация:

- соответствие компетенций планируемым результатам обучения по дисциплине (модулю);
- критерии оценивания сформированности компетенций;
- механизм формирования оценки по дисциплине (модулю);
- описание порядка применения и процедуры оценивания для каждого оценочного средства;
- критерии оценивания для каждого оценочного средства;
- содержание оценочных средств, включая требования, предъявляемые к действиям обучающихся, демонстрируемым результатам, задания различных типов.

Фонд оценочных средств по дисциплине находится в Приложении 1 к программе дисциплины (модулю).

7. Перечень литературы, необходимой для освоения дисциплины (модуля)

Освоение дисциплины (модуля) предполагает изучение основной и дополнительной учебной литературы. Литература может быть доступна обучающимся в одном из двух вариантов (либо в обоих из них):

- в электронном виде - через электронные библиотечные системы на основании заключенных КФУ договоров с правообладателями;
- в печатном виде - в Научной библиотеке им. Н.И. Лобачевского. Обучающиеся получают учебную литературу на абонементе по читательским билетам в соответствии с правилами пользования Научной библиотекой.

Электронные издания доступны дистанционно из любой точки при введении обучающимся своего логина и пароля от личного кабинета в системе "Электронный университет". При использовании печатных изданий библиотечный фонд должен быть укомплектован ими из расчета не менее 0,5 экземпляра (для обучающихся по ФГОС 3++ - не менее 0,25 экземпляра) каждого из изданий основной литературы и не менее 0,25 экземпляра дополнительной литературы на каждого обучающегося из числа лиц, одновременно осваивающих данную дисциплину.

Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля), находится в Приложении 2 к рабочей программе дисциплины. Он подлежит обновлению при изменении условий договоров КФУ с правообладателями электронных изданий и при изменении комплектования фондов Научной библиотеки КФУ.

8. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

Архив курсов ФКН ВШЭ - http://wiki.cs.hse.ru/Wiki_ФКН/Архив

Онлайн-платформа для дистанционного обучения stepik - https://stepik.org/

Платформа онлайн-обучения КФУ - https://edu.kpfu.ru

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Вид работ	Методические рекомендации
лекции	Студентам необходимо посещать лекции и вести конспект лекций вслед за изложением материала преподавателем. Рекомендуется прорабатывать конспект в течение дня после лекции и просматривать его вновь накануне следующей лекции. В случае обнаружения ошибок или возникновения вопросов по предыдущему материалу необходимо обратиться к преподавателю.

Вид работ	Методические рекомендации
практические занятия	Для подготовки к практическим занятиям студенту рекомендуется предварительно прорабатывать как лекционный материал, так и материал предыдущих практических занятий. Основой для подготовки служит добросовестное выполнение домашнего задания. Для успешного решения задач первой части курса студентам рекомендуется вспомнить материал, освоенный в предыдущих семестрах в рамках базовых математических дисциплин. Подготовку к семинарам (практическим занятиям, лабораторным занятиям) следует начинать с изучения теоретической части (лекционного материала) с определениями основных понятий, выводом формул и доказательством теорем. Особое внимание следует обращать на определения основных понятий и формулировки основных теорем. Необходимо подробно разбирать примеры, которые поясняют определения и теоремы. При разборе теорем необходимо учитывать, что все предположения теоремы должны использоваться в доказательстве ее утверждения, при этом необходимо понимать, в каком месте доказательства используется то или иное предположение теоремы. После изучения теоретического материала следует приступить к решениям задач по данной теме. Для многих задач курса существуют алгоритмы для их решения. В случае существования алгоритма решения задачи, необходимо разобрать все шаги работы этого алгоритма, обосновать, почему он останавливается через конечное число шагов и почему он дает необходимый результат. После этого при решении задач данного типа необходимо чётко следовать этому алгоритму.
самостоя-тельная работа	Самостоятельная работа студентов состоит из двух основных частей - проработка лекционного материала и выполнения домашних заданий. Для освоения теоретического и практического материала, в случае, когда конспектов оказывается недостаточным, или для более детальной проработки отдельных тем рекомендуется использовать литературу, указанную в соответствующем разделе. Все возникающие вопросы рекомендуется заранее четко сформулировать и впоследствии обсудить с преподавателем. Письменные домашние задания предназначены для самостоятельной проработки лекционного материала и овладения практическими навыками его применения для решения задач. Для освоения теоретического и практического материала, в случае, когда конспектов оказывается недостаточным, или для более детальной проработки отдельных тем рекомендуется использовать литературу, указанную в соответствующем разделе. Все возникающие вопросы рекомендуется заранее четко сформулировать и впоследствии обсудить с преподавателем.
экзамен	При подготовке к экзамену используйте литературу и источники, которые разбирались на семинарах в течение семестра. Ответ на экзамене предполагает полное и последовательное изложение изученного материала, а также демонстрацию способности и готовности применить полученные теоретические знания к предлагаемым практическим заданиям.

10. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения и информационных справочных систем (при необходимости)

Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения и информационных справочных систем, представлен в Приложении 3 к рабочей программе дисциплины (модуля).

11. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Материально-техническое обеспечение образовательного процесса по дисциплине (модулю) включает в себя следующие компоненты:

Помещения для самостоятельной работы обучающихся, укомплектованные специализированной мебелью (столы и стулья) и оснащенные компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду КФУ.

Учебные аудитории для контактной работы с преподавателем, укомплектованные специализированной мебелью (столы и стулья).

Компьютер и принтер для распечатки раздаточных материалов.

12. Средства адаптации преподавания дисциплины к потребностям обучающихся инвалидов и лиц с ограниченными возможностями здоровья

При необходимости в образовательном процессе применяются следующие методы и технологии, облегчающие восприятие информации обучающимися инвалидами и лицами с ограниченными возможностями здоровья:

- создание текстовой версии любого нетекстового контента для его возможного преобразования в альтернативные формы, удобные для различных пользователей;

- создание контента, который можно представить в различных видах без потери данных или структуры, предусмотреть возможность масштабирования текста и изображений без потери качества, предусмотреть доступность управления контентом с клавиатуры;

- создание возможностей для обучающихся воспринимать одну и ту же информацию из разных источников - например, так, чтобы лица с нарушениями слуха получали информацию визуально, с нарушениями зрения - аудиально;

- применение программных средств, обеспечивающих возможность освоения навыков и умений, формируемых дисциплиной, за счёт альтернативных способов, в том числе виртуальных лабораторий и симуляционных технологий;

- применение дистанционных образовательных технологий для передачи информации, организации различных форм интерактивной контактной работы обучающегося с преподавателем, в том числе вебинаров, которые могут быть использованы для проведения виртуальных лекций с возможностью взаимодействия всех участников дистанционного обучения, проведения семинаров, выступления с докладами и защиты выполненных работ, проведения тренингов, организации коллективной работы;

- применение дистанционных образовательных технологий для организации форм текущего и промежуточного контроля;

- увеличение продолжительности сдачи обучающимся инвалидом или лицом с ограниченными возможностями здоровья форм промежуточной аттестации по отношению к установленной продолжительности их сдачи:

- продолжительности сдачи зачёта или экзамена, проводимого в письменной форме, - не более чем на 90 минут;

- продолжительности подготовки обучающегося к ответу на зачёте или экзамене, проводимом в устной форме, - не более чем на 20 минут;

- продолжительности выступления обучающегося при защите курсовой работы - не более чем на 15 минут.

Программа составлена в соответствии с требованиями ФГОС ВО и учебным планом по направлению 01.04.01 "Математика" и магистерской программе "Анализ на многообразиях".

Приложение 2
к рабочей программе дисциплины (модуля)
Б1.В.ДВ.04.02 Введение в текстовую аналитику

Перечень литературы, необходимой для освоения дисциплины (модуля)

Направление подготовки: 01.04.01 - Математика
Профиль подготовки: Анализ на многообразиях
Квалификация выпускника: магистр
Форма обучения: очное
Язык обучения: русский
Год начала обучения по образовательной программе: 2025

Основная литература:

1.Батура, Т. В. Математическая лингвистика и автоматическая обработка текстов : учебное пособие / Батура Т. В. - Москва : Новосиб. гос. ун-т. - Ново-сибирск : РИЦ НГУ, 2016. - 166 с. - ISBN 978-5-4437-0548-4. - Текст : электронный // ЭБС 'Консультант студента' : [сайт]. - URL : https://www.studentlibrary.ru/book/ISBN9785443705484.html (дата обращения: 20.01.2025). - Режим доступа : по подписке.

2.Ингерсолл, Г. С. Обработка неструктурированных текстов. Поиск, организация и манипулирование / Г. С. Ингерсолл, Т. С. Мортон, Э. Л. Фэррис; пер. с англ. А. А. Слинкина. - 2-е изд. - Москва : ДМК Пресс, 2023. - 416 с. Систем. требования: Adobe Reader XI либо Adobe Digital Editions 4.5 ; экран 10'. - ISBN 978-5-89818-308-0. - Текст : электронный // ЭБС 'Консультант студента' : [сайт]. - URL : https://www.studentlibrary.ru/book/ISBN9785898183080.html (дата обращения: 20.01.2025). - Режим доступа : по подписке.

3. Коэльо, Л. П. Построение систем машинного обучения на языке Python / Л. П. Коэльо, В. Ричарт; пер. с англ. А. А. Слинкина. - 3-е изд. - Москва : ДМК Пресс, 2023. - 304 с. Систем. требования: Adobe Reader XI либо Adobe Digital Editions 4.5 ; экран 10'. - ISBN 978-5-89818-331-8. - Текст : электронный // ЭБС 'Консультант студента' : [сайт]. - URL : https://www.studentlibrary.ru/book/ISBN9785898183318.html (дата обращения: 20.01.2025). - Режим доступа : по подписке.

Дополнительная литература:

1. Риз, Р. Обработка естественного языка на Java / Р. Риз; пер. с англ. А. В. Снастина. - Москва : ДМК Пресс, 2016. - 264 с. - ISBN 978-5-97060-331-4. - Текст : электронный // ЭБС 'Консультант студента' : [сайт]. - URL : https://www.studentlibrary.ru/book/ISBN9785970603314.html (дата обращения: 20.01.2025). - Режим доступа : по подписке.

2. Йылдырым, С. Осваиваем архитектуру Transformer. Разработка современных моделей с помощью передовых методов обработки естественного языка / С. Йылдырым, М. Асгари-Ченаглу; пер. с анг. В. С. Яценкова. - Москва : ДМК Пресс, 2022. - 320 с. - ISBN 978-5-93700-106-1. - Текст : электронный // ЭБС 'Консультант студента' : [сайт]. - URL : https://www.studentlibrary.ru/book/ISBN9785937001061.html (дата обращения: 20.01.2025). - Режим доступа : по подписке.

3. Ферлитш, Э. Шаблоны и практика глубокого обучения / Э. Ферлитш; пер. с англ. А. В. Логунова. - Москва : ДМК Пресс, 2022. - 538 с. - ISBN 978-5-93700-113-9. - Текст : электронный // ЭБС 'Консультант студента' : [сайт]. - URL : https://www.studentlibrary.ru/book/ISBN9785937001139.html (дата обращения: 20.01.2025). - Режим доступа : по подписке.

Приложение 3
к рабочей программе дисциплины (модуля)
Б1.В.ДВ.04.02 Введение в текстовую аналитику

Перечень информационных технологий, используемых для освоения дисциплины (модуля), включая перечень программного обеспечения и информационных справочных систем

Направление подготовки: 01.04.01 - Математика

Профиль подготовки: Анализ на многообразиях

Квалификация выпускника: магистр

Форма обучения: очное

Язык обучения: русский

Год начала обучения по образовательной программе: 2025

Освоение дисциплины (модуля) предполагает использование следующего программного обеспечения и информационно-справочных систем:

Операционная система Microsoft Windows 7 Профессиональная или Windows XP (Volume License)

Пакет офисного программного обеспечения Microsoft Office 365 или Microsoft Office Professional plus 2010

Браузер Mozilla Firefox

Браузер Google Chrome

Adobe Reader XI или Adobe Acrobat Reader DC

Kaspersky Endpoint Security для Windows

Учебно-методическая литература для данной дисциплины имеется в наличии в электронно-библиотечной системе "ZNANIUM.COM", доступ к которой предоставлен обучающимся. ЭБС "ZNANIUM.COM" содержит произведения крупнейших российских учёных, руководителей государственных органов, преподавателей ведущих вузов страны, высококвалифицированных специалистов в различных сферах бизнеса. Фонд библиотеки сформирован с учетом всех изменений образовательных стандартов и включает учебники, учебные пособия, учебно-методические комплексы, монографии, авторефераты, диссертации, энциклопедии, словари и справочники, законодательно-нормативные документы, специальные периодические издания и издания, выпускаемые издательствами вузов. В настоящее время ЭБС ZNANIUM.COM соответствует всем требованиям федеральных государственных образовательных стандартов высшего образования (ФГОС ВО) нового поколения.

Учебно-методическая литература для данной дисциплины имеется в наличии в электронно-библиотечной системе "Консультант студента", доступ к которой предоставлен обучающимся. Многопрофильный образовательный ресурс "Консультант студента" является электронной библиотечной системой (ЭБС), предоставляющей доступ через сеть Интернет к учебной литературе и дополнительным материалам, приобретенным на основании прямых договоров с правообладателями. Полностью соответствует требованиям федеральных государственных образовательных стандартов высшего образования к комплектованию библиотек, в том числе электронных, в части формирования фондов основной и дополнительной литературы.