

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
"Казанский (Приволжский) федеральный университет"
Институт вычислительной математики и информационных технологий



подписано электронно-цифровой подписью

Программа дисциплины

Обработка естественного языка

Направление подготовки: 01.04.02 - Прикладная математика и информатика

Профиль подготовки: Анализ данных и его приложения

Квалификация выпускника: магистр

Форма обучения: очное

Язык обучения: русский

Год начала обучения по образовательной программе: 2025

Содержание

1. Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения ОПОП ВО
2. Место дисциплины (модуля) в структуре ОПОП ВО
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
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)
 - 4.2. Содержание дисциплины (модуля)
5. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине (модулю)
6. Фонд оценочных средств по дисциплине (модулю)
7. Перечень литературы, необходимой для освоения дисциплины (модуля)
8. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)
9. Методические указания для обучающихся по освоению дисциплины (модуля)
10. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения и информационных справочных систем (при необходимости)
11. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)
12. Средства адаптации преподавания дисциплины (модуля) к потребностям обучающихся инвалидов и лиц с ограниченными возможностями здоровья
13. Приложение №1. Фонд оценочных средств
14. Приложение №2. Перечень литературы, необходимой для освоения дисциплины (модуля)
15. Приложение №3. Перечень информационных технологий, используемых для освоения дисциплины (модуля), включая перечень программного обеспечения и информационных справочных систем

Программу дисциплины разработал(а)(и): старший преподаватель, к.н. Арабов М.К. (кафедра анализа данных и технологий программирования, отделение фундаментальной информатики и информационных технологий), MKArabov@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-4	Способен строить математические модели и анализировать данные, обосновывать и выбирать решение в прикладных задачах

Обучающийся, освоивший дисциплину (модуль):

Должен знать:

- базовые принципы обработки естественного языка;
- основные алгоритмы обработки естественного языка для решения задач машинного перевода и резюмирования текста.

Должен уметь:

- реализовывать основные алгоритмы обработки естественного языка в виде программы;
- применять алгоритмы обработки естественного языка для решения практических задач.

Должен владеть:

- терминологическим аппаратом обработки естественного языка
- навыками программной реализации алгоритмов машинного перевода и резюмирования текста

Должен демонстрировать способность и готовность:

- применять полученные знания и навыки в своей дальнейшей научно-исследовательской и профессиональной деятельности.

2. Место дисциплины (модуля) в структуре ОПОП ВО

Данная дисциплина (модуль) включена в раздел "Б1.В.ДВ.02.01 Дисциплины (модули)" основной профессиональной образовательной программы 01.04.02 "Прикладная математика и информатика (Анализ данных и его приложения)" и относится к дисциплинам по выбору части ОПОП ВО, формируемой участниками образовательных отношений.

Осваивается на 2 курсе в 3 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 2 зачетных(ые) единиц(ы) на 72 часа(ов).

Контактная работа - 32 часа(ов), в том числе лекции - 16 часа(ов), практические занятия - 0 часа(ов), лабораторные работы - 16 часа(ов), контроль самостоятельной работы - 0 часа(ов).

Самостоятельная работа - 40 часа(ов).

Контроль (зачёт / экзамен) - 0 часа(ов).

Форма промежуточного контроля дисциплины: зачет в 3 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	Лекции, всего	Лекции в эл. форме	Практические занятия, всего	Практические в эл. форме	Лабораторные работы, всего	Лабораторные в эл. форме	Самостоятельная работа
1.	Тема 1. Введение в обработку естественного языка	3	2	0	0	0	2	0	4
2.	Тема 2. Модель Bag-of-Words	3	2	0	0	0	2	0	6
3.	Тема 3. Векторные представления слов	3	2	0	0	0	2	0	6
4.	Тема 4. Рекуррентные нейронные сети в обработке естественного языка	3	2	0	0	0	2	0	6
5.	Тема 5. Контекстно-зависимые векторные представления слов	3	2	0	0	0	2	0	6
6.	Тема 6. Классификация текста	3	2	0	0	0	2	0	6
7.	Тема 7. Машинный перевод и резюмированное текста	3	2	0	0	0	2	0	6
4.2	Содержание дисциплины (модуля)								
	Тема 1. Введение в обработку естественного языка		4	0	0	0	4	0	6
	Задачи обработки естественного языка: категоризация текстов, выделение тем, машинный перевод, ответы на вопросы, анализ тональности текста, автоматическое реферирование и другие. Возможности методов машинного обучения в обработке естественного языка. История и развитие методов машинного обучения в обработке естественного языка. Приложения обработки естественного языка.								40

Тема 2. Модель Bag-of-Words

Составление словаря. Построение векторных представлений документов. Методы Term Frequency и Inverted Document Frequency как способы увеличения обобщающей способности моделей и эффективности. Предобработка данных: лемматизация, токенизация, исключение стоп-слов. Преимущества и ограничения модели Bag-of-words.

Тема 3. Векторные представления слов

Задача получения векторных представлений слов. Метод Word2Vec. Целевая функция и архитектура нейронной сети в алгоритме Word2Vec. Понятие контекста. Архитектуры CBOW и skip-gram в методе Word2Vec, их различия и ситуации применения. Повышение эффективности обучения нейронной сети на основе Word2Vec с помощью метода Negative Sampling. Свойства векторных представлений слов. Применение векторов Word2Vec.

Тема 4. Рекуррентные нейронные сети в обработке естественного языка

Общие принципы функционирования рекуррентных нейронных сетей. Виды ячеек в рекуррентных нейронных сетях, история развития, различие и эффективность наиболее популярных ячеек (LSTM, GRU). Архитектура LSTM-сетей. Применение рекуррентных сетей для построения языковой модели. Двухнаправленные и многослойные LSTM-сети.

Тема 5. Контекстно-зависимые векторные представления слов

Различия между контекстно-зависимым построением слов и обычным, сравнение с алгоритмом Word2Vec. Метод построения контекстно-зависимых векторных представлений слов. Обзор популярных методов построения контекстно-зависимых построений слов. Получение векторных представлений слов для задачи обучения с учителем на основе контекстно-зависимых векторов.

Тема 6. Классификация текста

Постановка задачи классификации текста. Приложения задачи классификации текста. Классификация текста с помощью рекуррентных нейронных сетей (LSTM-сетей). Классификация текста с помощью свёрточных нейронных сетей. Отличия подходов и сравнение их производительности. Метод классификации текста FastText.

Тема 7. Машинный перевод и резюмированное текста

Постановка задачи машинного перевода. Постановка задачи резюмирования текста. Понятие модели Sequence to Sequence, проблема генерации последовательностей разной длины. Архитектура нейронной сети Encoder-Decoder. Механизм внимания, принцип его работы и влияние на эффективность моделей. Применение моделей Sequence to Sequence и transformers

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 №14-55-99бин/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)

Интернет-портал образовательных ресурсов по ИТ - <http://www.intuit.ru>

Интернет-портал с научными ресурсами по математическим наукам - <http://www.mathnet.ru>

Цифровой образовательный ресурс "Нейронные сети и обработка текста" - <https://stepik.org/course/54098>

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

Вид работ	Методические рекомендации
лекции	Студенту рекомендуется внимательно слушать лектора, следить за тем, что написано на доске или представлено на слайдах презентации, анализировать получаемую им информацию. В случае, если материал лекции непонятен, следует задать вопрос в отведенное для вопросов время. Студенту также рекомендуется конспектировать материал лекции в тетради, что улучшает запоминание.
лабораторные работы	При выполнении лабораторных работ студенту рекомендуется внимательно анализировать поставленную задачу, уделяя особенное внимание критериям оценки точности решения задачи. Программный код должен быть объектно-ориентированным, чистым, с поясняющими комментариями. Особенное внимание следует уделить методологическим аспектам решения задачи, на корректное разделение выборки на обучающую, валидационную и тестовую. Результаты работы программы должны быть оформлены в виде таблиц и графиков.
самостоятельная работа	При ведении самостоятельной работы студенту рекомендуется внимательно подходить к изучению научных статей, обращать внимание на значимость полученного результата, на требования к обучающей выборке, на скорость работы предлагаемых алгоритмов, на результаты их сравнения с существующими. В случае, если изучаемый материал понятен не до конца, рекомендуется обращение к дополнительной литературе.
зачет	Студенту рекомендуется внимательно проанализировать вопросы в списке заданий. Ответ на задание должен быть подробным и четким, с приведением и объяснением всех релевантных формул. При ответе на вопрос студент должен продемонстрировать не только знание материала, но и его глубокое понимание. Следует избегать включения в ответ информации, не относящейся к заданию.

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 01.04.02 "Прикладная математика и информатика" и магистерской программе "Анализ данных и его приложения".

*Приложение 2
к рабочей программе дисциплины (модуля)
Б1.В.ДВ.02.01 Обработка естественного языка*

Перечень литературы, необходимой для освоения дисциплины (модуля)

Направление подготовки: 01.04.02 - Прикладная математика и информатика

Профиль подготовки: Анализ данных и его приложения

Квалификация выпускника: магистр

Форма обучения: очное

Язык обучения: русский

Год начала обучения по образовательной программе: 2025

Основная литература:

1. Гагарина, Л. Г. Введение в теорию алгоритмических языков и компиляторов : учебное пособие / Л.Г. Гагарина, Е.В. Кокорева. - 2-е изд., перераб. и доп. - Москва : ИНФРА-М, 2025. - 207 с. - (Высшее образование). - DOI 10.12737/1872635. - ISBN 978-5-16-017756-4. - Текст : электронный. - URL: <https://znanium.ru/catalog/product/1872635> (дата обращения: 10.01.2025). - Режим доступа: по подписке.
2. Саттон, Р. С. Обучение с подкреплением. Введение / Саттон Р. С., Барто Э. Дж. , пер. с англ. А. А. Слинкина. - 2-е изд. - Москва : ДМК Пресс, 2020. - 552 с. - ISBN 978-5-97060-097-9. - Текст : электронный // ЭБС 'Консультант студента' : [сайт]. - URL : <https://www.studentlibrary.ru/book/ISBN9785970600979.html> (дата обращения: 10.01.2025). - Режим доступа: по подписке.
3. Одинцов, Б. Е. Модели и проблемы интеллектуальных систем : монография / Б.Е. Одинцов. - Москва : ИНФРА-М, 2024. - 219 с. - (Научная мысль). - DOI 10.12737/1060845. - ISBN 978-5-16-015839-6. - Текст : электронный. - URL: <https://znanium.ru/catalog/product/2061202> (дата обращения: 10.01.2025). - Режим доступа: по подписке.

Дополнительная литература:

1. Обработка естественного языка с использованием языка программирования Python : учебное пособие : в 2 частях / составитель А. Б. Мантусов. - Элиста : КГУ, 2022 - Часть 1 - 2022. - 56 с. - Текст : электронный // Лань : электронно-библиотечная система. - URL: <https://e.lanbook.com/book/360923> (дата обращения: 10.01.2025). - Режим доступа: для авториз. пользователей.
2. Предобработка данных для нейросетевого управления : учебное пособие / Л. И. Воронова, В. Р. Брус, В. И. Воронов, А. Н. Баширов. - Москва : МТУСИ, 2021. - 49 с. - Текст : электронный // Лань : электронно-библиотечная система. - URL: <https://e.lanbook.com/book/215198> (дата обращения: 10.01.2025). - Режим доступа: для авториз. пользователей.
3. Корягин, С. В. Методы анализа естественно-языковых текстов : учебно-методическое пособие / С. В. Корягин, А. М. Русаков. - Москва : РТУ МИРЭА, 2023. - 86 с. - ISBN 978-5-7339-1737-5. - Текст : электронный // Лань : электронно-библиотечная система. - URL: <https://e.lanbook.com/book/331649> (дата обращения: 10.01.2025). - Режим доступа: для авториз. пользователей.

Приложение 3
к рабочей программе дисциплины (модуля)
Б1.В.ДВ.02.01 Обработка естественного языка

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

Направление подготовки: 01.04.02 - Прикладная математика и информатика

Профиль подготовки: Анализ данных и его приложения

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 Издательства "Лань", доступ к которой предоставлен обучающимся. ЭБС Издательства "Лань" включает в себя электронные версии книг издательства "Лань" и других ведущих издательств учебной литературы, а также электронные версии периодических изданий по естественным, техническим и гуманитарным наукам. ЭБС Издательства "Лань" обеспечивает доступ к научной, учебной литературе и научным периодическим изданиям по максимальному количеству профильных направлений с соблюдением всех авторских и смежных прав.

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.