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

УТВЕРЖДАЮ
Проректор по образовательной деятельности КФУ
_____ Турилова Е.А.
"___"_____20___ г.

Программа дисциплины

Обработка естественного языка

Направление подготовки: 01.04.02 - Прикладная математика и информатика
Профиль подготовки: Искусственный интеллект и суперкомпьютерные вычисления
Квалификация выпускника: магистр
Форма обучения: очное
Язык обучения: русский
Год начала обучения по образовательной программе: 2023

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): Першин И.А. ; заместитель директора по научной деятельности Тумаков Д.Н. (Директорат Института ВМ и ИТ, Институт вычислительной математики и информационных технологий), Dmitri.Tumakov@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	Разработка, отладка, рефакторинг программного кода, баз данных, информационных ресурсов; проектирование и интеграция программного обеспечения, управление проектами в области ИТ
ПК-5	Разработка, совершенствование и внедрение новых методов, технологий и инструментальных средств для получения, обработки, использования и аналитики больших данных.

Обучающийся, освоивший дисциплину (модуль):

Должен знать:

Линейную алгебру и аналитическую геометрию. Математическую статистику. Методы оптимизации. Машинное обучение. Язык программирования Python. Технологию параллельного программирования CUDA.

Должен уметь:

Считывать, предобрабатывать, визуализировать табличные данные.

Должен владеть:

Навыками параллельного программирования.

Должен демонстрировать способность и готовность:

Формировать математическую постановку решаемой задачи.

2. Место дисциплины (модуля) в структуре ОПОП ВО

Данная дисциплина (модуль) включена в раздел "Б1.В.06 Дисциплины (модули)" основной профессиональной образовательной программы 01.04.02 "Прикладная математика и информатика (Искусственный интеллект и суперкомпьютерные вычисления)" и относится к 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 2 зачетных(ые) единиц(ы) на 72 часа(ов).

Контактная работа - 18 часа(ов), в том числе лекции - 0 часа(ов), практические занятия - 0 часа(ов), лабораторные работы - 18 часа(ов), контроль самостоятельной работы - 0 часа(ов).

Самостоятельная работа - 54 часа(ов).

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

Форма промежуточного контроля дисциплины: зачет 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. Предварительная обработка текстовых данных.	2	0	0	0	0	6	0	18
2.	Тема 2. Рекуррентные нейронные сети. Архитектура Transformer.	2	0	0	0	0	6	0	18
3.	Тема 3. Архитектура BERT. Трансферное обучение.	2	0	0	0	0	6	0	18
	Итого		0	0	0	0	18	0	54

4.2 Содержание дисциплины (модуля)

Тема 1. Предварительная обработка текстовых данных.

Предварительная обработка текстовых данных - это процесс подготовки текстов для дальнейшего анализа или обработки с помощью алгоритмов машинного обучения. Она включает в себя различные шаги, такие как токенизация, удаление стоп-слов, приведение слов к нормальной форме, удаление пунктуации и специальных символов, а также преобразование текста в числовой формат, например, с помощью векторизации. Предварительная обработка текстовых данных помогает улучшить качество моделей машинного обучения, учитывая особенности текстов и устраняя шум или несущественную информацию. Это также может включать в себя методы для балансировки классов, обработки неизвестных слов или редких терминов, а также фильтрацию или аугментацию данных.

Тема 2. Рекуррентные нейронные сети. Архитектура Transformer.

Рекуррентные нейронные сети (англ. recurrent neural network, RNN) - это класс нейронных сетей, способных обрабатывать последовательности данных, такие как тексты или временные ряды. RNN применяются в задачах, таких как машинный перевод, генерация текста, распознавание речи и анализ тональности текста. Архитектура Transformer - это модель глубокого обучения, разработанная для обработки последовательностей данных, таких как тексты. Она основана на механизме внимания, который позволяет модели эффективно улавливать зависимости между различными элементами последовательности. В отличие от рекуррентных нейронных сетей, Transformer не требует последовательной обработки данных и может параллельно обрабатывать все элементы последовательности, что делает его более быстрым и эффективным.

Тема 3. Архитектура BERT. Трансферное обучение.

Архитектура BERT (Bidirectional Encoder Representations from Transformers) - это модель глубокого обучения, основанная на трансформерах, которая применяется для решения задач обработки естественного языка. Transfer learning - это метод обучения глубоких нейронных сетей, при котором модель предварительно обучается на большом наборе данных и затем переносится на новую задачу или набор данных, требующий меньшего объема обучающих примеров. Это позволяет значительно сократить время и ресурсы, необходимые для обучения модели, а также повысить ее производительность и точность на новой задаче.

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)

Классификация текста - https://lena-voita.github.io/nlp_course/text_classification.html

Модель Внимания - https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

Трансферное обучение - https://lena-voita.github.io/nlp_course/transfer_learning.html

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

Вид работ	Методические рекомендации
лабораторные работы	Перед началом работы необходимо ознакомиться с теоретическим материалом, связанным с темой лабораторной работы. Следует внимательно изучить постановку задачи и определить последовательность действий для ее выполнения. В процессе работы следует аккуратно фиксировать полученные результаты и делать соответствующие записи в лабораторном журнале. По окончании работы необходимо провести анализ полученных результатов, сделать выводы и подготовить отчет о выполненной лабораторной работе.
самостоя-тельная работа	При выполнении самостоятельной работы студент должен продемонстрировать навыки самостоятельного решения поставленной задачи. Для этого он должен предварительно ознакомиться с теоретическим материалом, Выбрать оптимальный метод решения поставленной задачи из списка допустимых методов, которые были рассмотрены на лекционных занятиях.

Вид работ	Методические рекомендации
зачет	В рамках зачета студент должен реализовать программу из фонда оценочных средств. Зачет проводится в компьютерной аудитории. Студент должен привести примеры использования приведенной программы, указать степень её важности. Если это возможно, то перечислить возможные способы модификации реализованной программы.

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 "Искусственный интеллект и суперкомпьютерные вычисления".

Приложение 2
к рабочей программе дисциплины (модуля)
Б1.В.06 Обработка естественного языка

Перечень литературы, необходимой для освоения дисциплины (модуля)

Направление подготовки: 01.04.02 - Прикладная математика и информатика
Профиль подготовки: Искусственный интеллект и суперкомпьютерные вычисления
Квалификация выпускника: магистр
Форма обучения: очное
Язык обучения: русский
Год начала обучения по образовательной программе: 2023

Основная литература:

1. Гафаров, Ф.М. Искусственные нейронные сети и приложения: учеб. пособие / Ф.М. Гафаров, А.Ф. Галимянов. - Казань: Изд-во Казан. ун-та, 2018. - 121 с. Режим доступа: https://kpfu.ru/staff_files/F1493580427/NejronGafGal.pdf

2. Применение искусственных нейронных сетей и системы остаточных классов в криптографии [Электронный ресурс] : монография / Н.И. Червяков [и др.]. ? Электрон. дан. ? Москва : Физматлит, 2012. ? 280 с. ? Режим доступа: https://e.lanbook.com/book/5300

3. Галушкин А.И., Нейронные сети: основы теории [Электронный ресурс] / Галушкин А.И. - М. : Горячая линия - Телеком, 2012. - 496 с. - ISBN 978-5-9912-0082-0 - Режим доступа: http://www.studentlibrary.ru/book/ISBN9785991200820.html

Дополнительная литература:

1. Тумаков, Д.Н. Технология программирования CUDA: учебное пособие / Д.Н. Тумаков, Д.Е. Чикрин, А.А. Егорчев, С.В. Голоусов. - Казань: Изд-во Казанского университета, 2017. - 112 с. (ISBN 978-5-00019-913-8) Режим доступа: https://shelly.kpfu.ru/e-ksu/docs/F244581044/Tekhnologiya_programmirovaniya_CUDA.pdf

2. Рутковская Д., Нейронные сети, генетические алгоритмы и нечеткие системы [Электронный ресурс] / Д. Рутковская, М. Пилиньский, Л. Рутковский. ; Пер. с польского И.Д. Рудинского. - 2-е изд., стереотип. - М. : Горячая линия - Телеком, 2012. - 384 с. - ISBN 978-5-9912-0320-3 - Режим доступа: http://www.studentlibrary.ru/book/ISBN9785991203203.html

3. Самко, Ю.Н. Психофизиология: Уч.пос. / Ю.Н. Самко. - М.: НИЦ ИНФРА-М, 2016. - 154 с. - (Высшее образование) ISBN 978-5-16-011402-6 - Режим доступа: http://znanium.com/catalog/product/520132

Приложение 3
к рабочей программе дисциплины (модуля)
Б1.В.06 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: Искусственный интеллект и суперкомпьютерные вычисления
Квалификация выпускника: магистр
Форма обучения: очное
Язык обучения: русский
Год начала обучения по образовательной программе: 2023

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

ЭЛЕКТРОННЫЙ УНИВЕРСИТЕТ
ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СИСТЕМА КФУ