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

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

Программа дисциплины

Основы обработки текстов на естественном языке

Направление подготовки: 09.04.04 - Программная инженерия
Профиль подготовки: Искусственный интеллект в разработке цифровых продуктов
Квалификация выпускника: магистр
Форма обучения: очное
Язык обучения: русский
Год начала обучения по образовательной программе: 2022

Содержание

ЭЛЕКТРОННЫЙ УНИВЕРСИТЕТ
ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СИСТЕМА КФУ

Программу дисциплины разработал(а)(и): доцент, к.н. (доцент) Липачев Е.К. (кафедра цифровой аналитики и технологий искусственного интеллекта., Институт информационных технологий и интеллектуальных систем), elipachev@gmail.com

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	Способность осуществлять проектирование и разработку интеллектуальных информационных систем
ПК-5	Способность выбирать наиболее подходящие модели и технологии искусственного интеллекта для решения задач профессиональной деятельности

Обучающийся, освоивший дисциплину (модуль):

Должен знать:

Должен знать:

Основные этапы обработки естественно-языковых текстов, какие задачи возникают на каждом этапе и основные подходы к их решению,

- знать об основных подходах проверки качества методов решения основных типов задач,

- знать теорию вероятности, дискретную математику и линейную алгебру на уровне бакалавриата ИТИС.

Должен уметь:

Должен уметь:

- применять существующие программные библиотеки для решения задач обработки текста (Apache UIMA, GATE, Apache OpenNLP),

- уметь понимать принципы представления текстовых данных в ЭВМ, применяемых для их хранения и передачи по каналам связи,

- уметь проектировать и разрабатывать программные приложения, осуществляющие интеллектуальную обработку входных текстовых данных.

Должен владеть:

Должен владеть:

- владеть основами программирования,

- владеть языком программирования (Java, Python).

2. Место дисциплины (модуля) в структуре ОПОП ВО

Данная дисциплина (модуль) включена в раздел "Б1.В.ДВ.03.01 Дисциплины (модули)" основной профессиональной образовательной программы 09.04.04 "Программная инженерия (Искусственный интеллект в разработке цифровых продуктов)" 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 4 зачетных(ые) единиц(ы) на 144 часа(ов).

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

Самостоятельная работа - 90 часа(ов).

Контроль (зачёт / экзамен) - 18 часа(ов).

ЭЛЕКТРОННЫЙ УНИВЕРСИТЕТ
ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СИСТЕМА КФУ

Форма промежуточного контроля дисциплины: экзамен 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	
1.	Тема 1. Тема 1. Языковые модели. Основы обработки текстов на естественном языке словарными подходами и методами, основанными на правилах.	3	0	0	0	0	10	0	25
2.	Тема 2. Тема 2. Обработка текстов на естественном языке методами машинного обучения.	3	0	0	0	0	14	0	35
3.	Тема 3. Тема 3. Прикладные задачи обработки текстов на естественном языке.	3	0	0	0	0	12	0	30
	Итого		0	0	0	0	36	0	90

4.2 Содержание дисциплины (модуля)

Тема 1. Тема 1. Языковые модели. Основы обработки текстов на естественном языке словарными подходами и методами, основанными на правилах.

Языковые модели. Анализ текста в парадигме когнитивных исследований. Анализ текста в парадигмах автоматического понимания текста. Коммуникативная и информационная (смысловая) структуры текста.

Избыточность. Морфологический анализ и синтез. Словарный морфологический анализ и синтез. Компрессия текста.

Лемматизация. Приведение текста в нормальную форму. Классификация текста. Оценка качества классификации. Создание словарей. Методы разработки правил.

Тема 2. Тема 2. Обработка текстов на естественном языке методами машинного обучения.

Наивный Байесовский классификатор. Теорема Байеса. Функции ошибки и регуляризация. KL-расстояние и перекрестная энтропия. Градиентный спуск: основы. Граф вычислений и дифференцирование на нем.

Логистическая регрессия. Метод опорных векторов. Деревья решений.

Автоматическое извлечение признаков из текста. Подбор оптимальных признаков.

Тема 3. Тема 3. Прикладные задачи обработки текстов на естественном языке.

Извлечение информации из текстов, выделение именованных сущностей.

Анализ тональности текстов. Задача поиска отношений.

Разрешение анафоры и кореференции.

Машинный перевод.

Синтаксические парсеры русского и английского языка.

Вопросно-ответные системы. Диалоговые системы и чат-боты.

Онтологии и тезаурусы.

Тематическое моделирование.

Кластеризация текстов.

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

ЭЛЕКТРОННЫЙ УНИВЕРСИТЕТ
ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СИСТЕМА КФУ

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)

Каталог ресурсов для обработки естественного языка. - https://nlpub.ru/

YSDA Natural Language Processing course. - https://github.com/yandexdataschool/nlp_course/tree/master.

ЭЛЕКТРОННЫЙ УНИВЕРСИТЕТ
ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СИСТЕМА КФУ

Техносфера. Информационный поиск. Видеолекции. - https://sphere.mail.ru/materials/video/

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

Вид работ	Методические рекомендации
лабораторные работы	Должна решать одну из поставленных задач по обработке текстов на естественном языке. При выполнении практического задания рекомендуется пользоваться лекционным материалом, а также навыками, приобретенными на занятиях. Программный код должен быть хорошо структурирован, оформлен согласно правилам, установленным для языка программирования и выполняться без ошибок.
самостоя-тельная работа	Самостоятельная работа включает в себя выполнение домашней работы в течение семестра, выполнение курсовой работы, подготовку научного доклада, а также чтение дополнительной литературы при необходимости. Работу над курсовой рекомендуется начать сразу после получения задачи, поскольку расчеты могут занять длительное время.
экзамен	В качестве материалов при подготовке к экзамену рекомендуется использовать презентации и рекомендуемую литературу. Ответ на вопрос должен включать в себя: формулировку постановки задачи, описание методов решения и методов оценки качества. Если в вопросе содержится описание метода, необходимо описать его математическую модель формулами. Необходимо дать максимально полный развернутый ответ на вопрос.

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;

ЭЛЕКТРОННЫЙ УНИВЕРСИТЕТ
ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СИСТЕМА КФУ

- 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 09.04.04 "Программная инженерия" и магистерской программе "Искусственный интеллект в разработке цифровых продуктов".

ЭЛЕКТРОННЫЙ УНИВЕРСИТЕТ
ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СИСТЕМА КФУ

Приложение 2
к рабочей программе дисциплины (модуля)
Б1.В.ДВ.03.01 Основы обработки текстов на естественном языке

Перечень литературы, необходимой для освоения дисциплины (модуля)

Направление подготовки: 09.04.04 - Программная инженерия

Профиль подготовки: Искусственный интеллект в разработке цифровых продуктов

Квалификация выпускника: магистр

Форма обучения: очное

Язык обучения: русский

Год начала обучения по образовательной программе: 2022

Основная литература:

1.Городняя, Л. В. Парадигма программирования : учебное пособие для вузов / Л. В. Городняя. - 2-е изд., стер. - Санкт-Петербург : Лань, 2021. - 232 с. - ISBN 978-5-8114-6680-1. - Текст : электронный // Лань : электронно-библиотечная система. - URL: https://e.lanbook.com/book/151660 (дата обращения: 21.02.2022). - Режим доступа: для авториз. пользователей.

2. Хиценко В.П., Основы программирования : учебное пособие / Хиценко В.П. - Новосибирск : Издательство Новосибирского государственного технического университета, 2015. - 83 с. - ISBN 978-5-7782-2706-4 - Текст : электронный // ЭБС 'Консультант студента' : [сайт]. - URL : http://www.studentlibrary.ru/book/ISBN9785778227064.html (дата обращения: 21.02.2022). - Режим доступа : по подписке.

3. Грант С.И., Обработка неструктурированных текстов. Поиск, организация и манипулирование: монография / Грант С. Ингерсолл, Томас С. Мортон, Эндрю Л. Фэррис - Москва: ДМК Пресс, 2015. - 414 с. - ISBN 978-5-97060-144-0 - Текст : электронный // ЭБС 'Консультант студента' : [сайт]. - URL : http://www.studentlibrary.ru/book/ISBN9785970601440.html (дата обращения: 21.02.2022). - Режим доступа : по подписке

Дополнительная литература:

1.Батура Т.В., Математическая лингвистика и автоматическая обработка текстов : учебное пособие / Батура Т.В. - Новосибирск : РИЦ НГУ, 2016. - 166 с. - ISBN 978-5-4437-0548-4 - Текст : электронный // ЭБС 'Консультант студента' : [сайт]. - URL : http://www.studentlibrary.ru/book/ISBN9785443705484.html (дата обращения: 21.02.2022). - Режим доступа : по подписке.

2. Хожемпо В.В., Азбука научно-исследовательской работы студента : учебное пособие / В.В. Хожемпо, К.С. Тарасов, М.Е. Пухлянко. - изд. 2-е, испр. и доп. - Москва: Издательство РУДН, 2010. - 107 с. - ISBN 978-5-209-03527-5 - Текст : электронный // ЭБС 'Консультант студента' : [сайт]. - URL : http://www.studentlibrary.ru/book/ISBN9785209035275.html (дата обращения: 21.02.2022). - Режим доступа : по подписке.

3.Лукашевич Н.В., Тезаурусы в задачах информационного поиска: монография / Лукашевич Н.В. - Москва: Издательство Московского государственного университета, 2011. - 512 с. - ISBN 978-5-211-05926-9 - Текст : электронный // ЭБС 'Консультант студента' : [сайт]. - URL : http://www.studentlibrary.ru/book/ISBN9785211059269.html (дата обращения: 21.02.2022). - Режим доступа : по подписке.

ЭЛЕКТРОННЫЙ УНИВЕРСИТЕТ
ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СИСТЕМА КФУ

Приложение 3
к рабочей программе дисциплины (модуля)
Б1.В.ДВ.03.01 Основы обработки текстов на естественном языке

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

Направление подготовки: 09.04.04 - Программная инженерия

Профиль подготовки: Искусственный интеллект в разработке цифровых продуктов

Квалификация выпускника: магистр

Форма обучения: очное

Язык обучения: русский

Год начала обучения по образовательной программе: 2022

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 "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.

ЭЛЕКТРОННЫЙ УНИВЕРСИТЕТ
ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СИСТЕМА КФУ