МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное бюджетное образовательное учреждение высшего образования

«ИРКУТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

(ФГБОУ ВО «ИГУ»)

Институт математики и информационных технологий

Кафедра алгебраических и информационных систем

«УТВЕРЖДАЮ»

Директор ИМИТ ИГУ

_____ М. В. Фалалеев

«15» мая 2024 г.

Рабочая программа дисциплины (модуля)

Б1.В.ДВ.01.02 Обработка текстов на естественном языке

Направление подготовки 02.04.02 Фундаментальная информатика и информационные технологии

Направленность (профиль) подготовки Анализ данных научных исследований и машинное обучение

Квалификация выпускника магистр

Форма обучения очная

Иркутск 2024 г.

1. Цели и задачи дисциплины:

Цель

Целью дисциплины Б1.В.ДВ.01.02 **Обработка текстов на естественном языке** является овладение студентами навыков применения и математических методов и алгоритмов для задач обработки и анализа текстов на естественном языке.

Задачи:

Познакомить студентов с основными задачами обработки текстов, с моделями текстов, с алгоритмами преобразования и обработки текстов, а также с распространенными библиотеками, которые используются при решении задач обработки текстов.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП ВО

2.1. Учебная дисциплина (модуль) Б1.В.ДВ.01.02 **Обработка текстов на естественном языке** относится к вариативной части программы.

2.2. Для изучения данной учебной дисциплины (модуля) необходимы знания, умения и навыки, формируемые дисциплинами, включенными в программу бакалавриата: высшая математика, математический анализ, дискретная математика, линейная алгебра, теория вероятностей. В программе магистратуры к предшествующим дисциплинам относятся: прикладная статистика, математика для анализа данных, анализ и визуализация данных, машинное обучение, глубокое обучение.

2.3. Знания, умения и навыки, формируемые данной учебной дисциплиной, могут быть использованы при написании выпускной квалификационной работы а также в процессе научно-исследовательской работы.

3. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Процесс освоения дисциплины направлен на формирование компетенций (элементов следующих компетенций) в соответствии с ФГОС ВО по соответствующему направлению подготовки.

Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с индикаторами достижения компетенций

Компетенция	Индикаторы компетенций	Результаты обучения
ПК-3. Способен формулировать задачи, анализировать и применять способы и методы научных исследований, проводить информационный поиск и	ИДК $_{ОПК1.1}$ Умеет выделять проблемы, относящиеся к прикладной математике, фундаментальной информатике и информационным технологиям	Знать: основные понятия и методы обработки текстовой информации, техники предобработки текстов, методы интеллектуального анализа текстов на естественном языке Уметь: использовать математические методы и модели

использовать информационные ресурсы для решения научно-исследовательских задач, формулировать и представлять научные результаты в форме презентаций и публикаций	ИДК ОПК1.2 Умеет решать актуальные проблемы прикладной математики, фундаментальной информатики и информационных технологий	машинного обучения для обработки текстов на естественном языке. Владеть: навыками применения языка программирования Python для решения задач обработки текстов на естественном языке.
	ИДК ОПК1.3 Способен формулировать проблемы прикладной математики, фундаментальной информатики и информационных технологий	

4. СОДЕРЖАНИЕ И СТРУКТУРА ДИСЦИПЛИНЫ

Трудоемкость дисциплины составляет 2 зачетных единицы, 72 часа,

Форма промежуточной аттестации: зачет

4.1 Содержание дисциплины, структурированное по темам, с указанием видов учебных занятий и отведенного на них количества академических часов

4.2 План внеаудиторной самостоятельной работы обучающихся по дисциплине

Семестр	Название раздела, темы	Самостоятельная работа обучающихся			Оценочное средство	Учебно-методическое обеспечение самостоятельной работы
		Вид самостоятельной работы	Сроки выполнения	Затраты времени (час.)		
1	**Предобработка текста**:	*УИЛК*	1-я половина курса	18	контроль выполнения работ	Лабораторные работы в ИОС Домик
	Регулярные выражения. Использование регулярных выражений в Python.	*УИЛК*		2		
	Токенизация с использованием регулярных выражений.	*УИЛК*		4		
	Проверка орфографии. Редакторское расстояние.	*УИЛК*		4		
	Стемминг и лемматизация.	*УИЛК*		4		
	КС-грамматики. Построение дерева разбора.	*УИЛК*		4		
	Понимание текста:	*УИЛК*	2-я половина курса	10	контроль выполнения работ	Лабораторные работы в ИОС Домик
	Модель мешка слов.	*УИЛК*		2		
	Модель Word2Vec.	*УИЛК*		4		
	Семантические сети.	*УИЛК*		4		
	Использование методов машинного обучения для решения задач обработки текстов:	*УИЛК*	2-я половина курса	12	контроль выполнения работ	Лабораторные работы в ИОС Домик
	Логистическая регрессия	*УИЛК*		6		
	Деревья решений и случайный лес	*УИЛК*		6		
Общая трудоемкость самостоятельной работы по дисциплине (час)				**40**		
Бюджет времени самостоятельной работы, предусмотренный учебным планом для данной дисциплины (час)				**40**		

Виды самостоятельной работы:
Р – написание реферата,
Д – подготовка доклада,

У – выполнение упражнений,
Э – написание эссе,
Пт – выполнение проекта,
К - кейс-задание,
Пф – портфолио,
И – информационный поиск,
Прз – презентация,
Л – изучение литературы,
Т – заполнение таблицы Донны Огл «Знал, хотел узнать, узнал»
Ин – заполнение таблицы, содержащей 4 столбца – «V» - уже знал, «+» - новое, «-» – думал иначе, «?» – не понял, есть вопросы.

4.3. Методические указания по организации самостоятельной работы студентов

Методические указания по организации самостоятельной работы расположены в ИОС Educa

5. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

а) основная литература

1. Введение в информационный поиск [Текст] : научное издание / К. Д. Маннинг, П. Рагхаван, Х. Шютце ; пер. с англ. Д. А. Клюшин. - М. : Вильямс, 2014. - 520 с. ; 24 см. - Библиогр.: с. 473-505. - ISBN 978-5-8459-1623-5.

2. Коэльо, Л. П. Построение систем машинного обучения на языке Python / Л. П. Коэльо, В. Ричарт ; перевод с английского А. А. Слинкин. — 2-е изд. — Москва : ДМК Пресс, 2016. — 302 с. — ISBN 978-5-97060-330-7. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/82818 (дата обращения: 30.06.2021). — Режим доступа: для авториз. пользователей.

Дополнительная литература

1. Антонио, Д. Библиотека Keras – инструмент глубокого обучения. Реализация нейронных сетей с помощью библиотек Theano и TensorFlow / Д. Антонио, П. Суджит ; перевод с английского А. А. Слинкин. — Москва : ДМК Пресс, 2018. — 294 с. — ISBN 978-5-97060-573-8. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/111438 (дата обращения: 30.06.2021). — Режим доступа: для авториз. пользователей.

в) периодические издания

1. CODATA Data Science Journal https://datascience.codata.org/ CODATA Data Science Journal - рецензируемый электронный журнал с открытым доступом, публикующий статьи по управлению, распространению, использованию и повторному использованию исследовательских данных и баз данных во всех областях исследований, включая науку, технологии, гуманитарные науки и искусство.

г) список авторских методических разработок: *(Указываются при наличии. Если имеются, то указываются учебники, учебные пособия, авторские лекции, методические рекомендации, программы и др.включая информацию о материалах размещенных в ЭИОС ИГУ(КДО))*

д) базы данных, информационно-справочные и поисковые системы

1. Единое окно доступа к образовательным ресурсам. Полнотекстовая электронная библиотека учебных и учебно-методических материалов (федеральный ресурс).

http://www.window.edu.ru.

2. Электронно-библиотечная система издательства «ЮРАЙТ» https://www.biblio-online.ru/

3. Электронно-библиотечная система издательства «Лань» https://e.lanbook.com/

4. ИОС ИГУ EDuca

5. Образовательная платформа https://stepik.org

6. Образовательный онлайн-проект https://www.coursera.org/

7. Indigo — внедрение data science-решений на примере конкретных кейсов, рассказанных компанией-разработчиком.

8. Google Colaboratory https://colab.research.google.com

9. Your machine learning and data science community https://kaggle.com.

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1. Учебно-лабораторное оборудование:

Для проведения лекционных занятий необходима аудитория с презентационным оборудованием, для проведения практических занятий необходима компьютерная аудитория на 10-20 рабочих мест (в зависимости от численности учебной группы), оборудованная доской, презентационной техникой.

6.2. Программное обеспечение:

Среда программирования Python, браузер

6.3. Технические и электронные средства:

ИОС EDUCA, DOMIC, презентационное оборудование, персональный компьютер с возможностью просмотра презентаций.

7. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

При реализации данного курса используются следующие образовательные технологии: технологии традиционного обучения, игровые технологии, технологии проблемного обучения, технологии обучения в сотрудничестве, технологии контекстного обучения, интерактивные технологии, технологии дистанционного обучения, активные педагогические технологии.

8. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

Оценочные средства (ОС):

8.1. Входной контроль не предусмотрен

8.2. Оценочные средства текущего контроля – проверка выполнения лабораторных работ.

8.3. Оценочные средства для промежуточной аттестации (в форме зачета).

Зачет выставляется по баллам, полученным за выполнение лабораторных работ в соответствии с балльно-рейтинговой системой университета. Для получения зачета необходимо набрать не менее 60 баллов. Выполнение каждой лабораторной работы оценивается в 10 баллов.

Список лабораторных работ.

1. Регулярные выражения. Использование регулярных выражений в Python.
2. Токенизация с использованием регулярных выражений.
3. Проверка орфографии. Редакторское расстояние.
4. Стемминг и лемматизация.
5. КС-грамматики. Построение дерева разбора.
6. Модель мешка слов.
7. Модель Word2Vec.
8. Семантические сети.
9. Логистическая регрессия
10. Деревья решений и случайный лес

Разработчики:

доцент	Кириченко Константин Дмитриевич	
(подпись)	(занимаемая должность)	(Ф.И.О.)

Программа составлена в соответствии с требованиями ФГОС ВО по направлению подготовки 02.04.02 «Фундаментальная информатика и информационные технологии» (уровень магистратуры), утвержденный приказом Министерства образования и науки Российской Федерации от «23» августа 2017 г. № 811, зарегистрированный в Минюсте России «13» сентября 2017 г. № 48168 с изменениями и дополнениями от 26 ноября 2020 г., 8 февраля 2021 г.

Программа рассмотрена на заседании кафедры Алгебраических и информационных систем ИМИТ ИГУ «24» марта 2022 г.

Протокол № 9 Зав. кафедрой_____ Пантелеев В.И.

Настоящая программа, не может быть воспроизведена ни в какой форме без предварительного письменного разрешения кафедры-разработчика программы.