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

«УТВЕРЖДАЮ»
Директор ИМИТ ИГУ
_____ М. В. Фалалеев
«10» мая 2023 г.

Рабочая программа дисциплины (модуля)

Б1.О.10 Обработка естественных языков

Направление подготовки: 01.04.02 Прикладная математика и информатика

Направленность (профиль) подготовки: Семантические технологии и многоагентные системы

Квалификация выпускника: магистр

Форма обучения: очная

Иркутск 2023 г.

2 АННОТАЦИЯ ДИСЦИПЛИНЫ

«ОБРАБОТКА ЕСТЕСТВЕННЫХ ЯЗЫКОВ»

В курсе рассматриваются задачи, которые требуют обработки текстов на естественных языках, в первую очередь русском и английском. Список задач включает в себя классификацию текстов, определение тональности, автоматическое реферирование, машинный перевод, многие другие задачи более низкого уровня. Из подходов к решению задач рассматриваются лингвистические подходы, статистические и подходы, использующие глубокое обучение. Курс предполагает решение практических заданий с помощью библиотек и ресурсов для обработки естественных языков для языка программирования python.

SUBJECT SUMMARY

«NATURAL LANGUAGE PROCESSING»

The course is devoted to the approaches to solve problems, that require processing of raw texts in natural languages, primarily Russian and English. The list of problems includes texts classification, sentiment analysis, automatic summarization, machine translation, a number of other low level tasks. The approaches being discussed include purely linguistic approaches, statistical approaches and deep learning approaches. The course supposes practice in coding in python using natural language processing libraries and resources.

3 ОБЩИЕ ПОЛОЖЕНИЯ

3.1 Цели и задачи дисциплины

1. Цель дисциплины -изучение математического аппарата, используемого в основе методов обработки естественных языков, и программных инструментов для обработки естественных языков и приобретение практических навыков в профессиональной деятельности.

2. Изучение прикладных проблем, которые возникают при обработке текстов на естественных языках и подходы к их решению. Освоение классификации текстов, определение тональности, автоматическое реферирование, машинный перевод, многие другие задачи более низкого уровня. Приобретение навыков решения практических заданий с помощью библиотек и ресурсов для обработки естественных языков для языка программирования python.

3. Знания спектра подходов к решению разных задач обработки естественных языков, методов обработки естественных языков, программных инструментов, используемых при различных подходах и методах.

4. Умения связать решаемую задачу с задачами из области обработки естественных языков. Умение выбрать подход к решению задачи.

5. Формирование навыков использования математического аппарата и программных инструментов для решения проблематики обработки естественных языков.

3.2 Место дисциплины в структуре ОПОП

Дисциплина изучается на основе ранее освоенных дисциплин учебного плана:

1. «Статистика случайных процессов»

2. «Интеллектуальные системы»

3. «Алгоритмы компьютерной математики»

4. «Машинное обучение»

5. «Семантический Web»

и обеспечивает изучение последующих дисциплин:

1. «Производственная практика (преддипломная практика)»

3.3 Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы

В результате освоения образовательной программы обучающийся должен достичь следующие результаты обучения по дисциплине:

Код компетенции/ индикатора компетенции	Наименование компетенции/индикатора компетенции
ОПК-2	Способен совершенствовать и реализовывать новые математические методы решения прикладных задач
ОПК-2.1	*Знает современные математические методы решения прикладных задач*
ОПК-2.2	*Умеет обосновывать выбор либо необходимость реализации новых математических методов решения прикладных задач*
ОПК-2.3	*Знает принципы и основные современные методы решения задач управления в технических системах*

4 СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

4.1 Содержание разделов дисциплины

4.1.1 Наименование тем и часы на все виды нагрузки

№ п/п	Наименование темы дисциплины	Лек, ач	Пр, ач	ИКР, ач	СР, ач
1	Введение	2	2		12
2	Задача классификации	2	2		10
3	Работа с последовательностями	2	2		10
4	Теория формальных языков в приложении к естественным	2	2		10
5	Семантика	2	2		18
6	Трансформеры последовательностей	3	3		10
7	Заключение	1	1		1
	Итого, ач	14	14	10	71

	Из них ач на контроль	0	0	0	35
	Общая трудоемкость освоения, ач/зе	144/4			

4.1.2 Содержание

№ п/п	Наименование темы дисциплины	Содержание
1	Введение	Задачи обработки естественного языка, этапы обработки текста.
2	Задача классификации	Наивный байесовый классификатор и другие традиционные методы классификации. Классификация с помощью нейронных сетей прямого распространения. Лингвистические приложения методов классификации.
3	Работа с последовательностями	Текст как последовательность символов или слов. N-грамм модели, модели основанные на рекуррентных нейронных сетях. Сглаживание, оценка модели. Классификация элементов последовательности.
4	Теория формальных языков в приложении к естественным	Контекстно-свободные грамматики в описании естественных языков, дерево разбора текста и граф зависимостей.
5	Семантика	Разные подходы к определению понятия смысла языковых единиц: исчисление предикатов, дистрибутивная семантика. Кластеризация. Нейросетевые подходы к векторному представлению слов (word embedding). Разрешение кореферентности.
6	Трансформеры последовательностей	Понятие трансформера, энкодера, декодера, механизм внимания, современные нейросетевые модели. Машинный перевод, генерация текста.
7	Заключение	Обзор других современных направлений в обработке естественных языков.

4.2 Перечень лабораторных работ

Лабораторные работы не предусмотрены.

4.3 Перечень практических занятий

Наименование практических занятий	Количество ауд. часов
1. Определение авторства текста	1
2. N-грамм модели	1
3. Определение частей речи в тексте	1
4. Грамматический разбор текста	1
5. Поиск коллокаций	2
6. Word2Vec, кластеризация слов	2

7. BERT для анализа настроения	2
8. Выделение именованных сущностей	2
9. GPT2 для генерации текста	2
Итого	14

4.4 Курсовое проектирование

Курсовая работа (проект) не предусмотрены.

4.5 Реферат

Реферат не предусмотрен.

4.6 Индивидуальное домашнее задание

Индивидуальное домашнее задание не предусмотрено.

4.7 Доклад

Доклад не предусмотрен.

4.8 Кейс

Кейс не предусмотрен.

4.9 Организация и учебно-методическое обеспечение самостоятельной работы

Изучение дисциплины сопровождается самостоятельной работой студентов с рекомендованными преподавателем литературными источниками и информационными ресурсами сети Интернет.

Планирование времени для изучения дисциплины осуществляется на весь период обучения, предусматривая при этом регулярное повторение пройденного материала. Обучающимся, в рамках внеаудиторной самостоятельной работы, необходимо регулярно дополнять сведениями из литературных источников

материал, законспектированный на лекциях. При этом на основе изучения рекомендованной литературы целесообразно составить конспект основных положений, терминов и определений, необходимых для освоения разделов учебной дисциплины.

Особое место уделяется консультированию, как одной из форм обучения и контроля самостоятельной работы. Консультирование предполагает особым образом организованное взаимодействие между преподавателем и студентами, при этом предполагается, что консультант либо знает готовое решение, которое он может предписать консультируемому, либо он владеет способами деятельности, которые указывают путь решения проблемы.

Текущая СРС	**Примерная трудоемкость, ач**
Работа с лекционным материалом, с учебной литературой	34
Опережающая самостоятельная работа (изучение нового материала до его изложения на занятиях)	0
Самостоятельное изучение разделов дисциплины	0
Выполнение домашних заданий, домашних контрольных работ	20
Подготовка к лабораторным работам, к практическим и семинарским занятиям	17
Подготовка к контрольным работам, коллоквиумам	0
Выполнение расчетно-графических работ	0
Выполнение курсового проекта или курсовой работы	0
Поиск, изучение и презентация информации по заданной проблеме, анализ научных публикаций по заданной теме	0
Работа над междисциплинарным проектом	0
Анализ данных по заданной теме, выполнение расчетов, составление схем и моделей, на основе собранных данных	0
Подготовка к зачету, дифференцированному зачету, экзамену	35
ИТОГО СРС	106

5 Учебно-методическое обеспечение дисциплины

5.1 Перечень основной и дополнительной литературы, необходимой для освоения дисциплины

№ п/п	**Название, библиографическое описание**	**К-во экз. в библ.**
Основная литература		

1	Гольдберг Й. Нейросетевые методы в обработке естественного языка [Электронный ресурс] : руководство / Й. Гольдберг, 2019. -282 с.	неогр.
2	Лейн Хобсон Обработка естественного языка в действии [Электронный ресурс] / Хобсон Лейн, Ханнес Хапке, Коул Ховард, 2021. -576 с.	неогр.
Дополнительная литература		
1	Ганегедара Т. Обработка естественного языка с TensorFlow [Электронный ресурс] : руководство / Т. Ганегедара, 2020. -382 с.	неогр.
2	Риз Р. Обработка естественного языка на Java [Электронный ресурс], 2016. -264 с.	неогр.

5.2 Перечень ресурсов информационно-телекоммуникационной сети «Интернет», используемых при освоении дисциплины

№ п/п	Электронный адрес
1	Сервис NLPub по созданию и развитию русских языковых ресурсов: https://nlpub.ru/
2	Проект «Открытый корпус» создания морфологически, синтаксически и семантически размеченных корпусов текстов на русском языке: http://opencorpora.org/
3	Natural Language Toolkit: https://www.nltk.org/
4	Natural Language Processing, Jacob Eisenstein: https://raw.githubusercontent.com/jacobeisnlp-class/master/notes/eisenstein-nlp-notes.pdf
5	Mining of Massive Datasets: http://mmds.org/

6 Критерии оценивания и оценочные материалы

6.1 Критерии оценивания

Для дисциплины «Обработка естественных языков» формой промежуточной аттестации является экзамен.

Экзамен

Оценка	Описание
Неудовлетворительно	Курс не освоен. Студент испытывает серьезные трудности при ответе на ключевые вопросы дисциплины.
Удовлетворительно	Студент в целом овладел курсом, но некоторые разделы освоены на уровне определений и формулировок теорем.
Хорошо	Студент овладел курсом, но в отдельных вопросах испытывает затруднения. Умеет решать задачи.
Отлично	Студент демонстрирует полное овладение курсом, способен применять полученные знания при решении конкретных задач.

Особенности допуска

По результатам текущего контроля (выполнения всех параметров более чем на 60 % (баллов)) студент получает допуск на экзамен.

6.2 Оценочные материалы для проведения текущего контроля и промежуточной аттестации обучающихся по дисциплине

Примерные вопросы к экзамену

№ п/п	Описание
1	Задачи обработки естественного языка, этапы обработки текста.
2	Наивный байесовый классификатор и другие традиционные методы классификации.
3	Классификация с помощью нейронных сетей прямого распространения.
4	Лингвистические приложения методов классификации.
5	Текст как последовательность символов или слов.
6	N-грамм модели, модели основанные на рекуррентных нейронных сетях.
7	Сглаживание, оценка модели.
8	Классификация элементов последовательности символов и слов.
9	Контекстно-свободные грамматики в описании естественных языков, дерево разбора текста.
10	Контекстно-свободные грамматики в описании естественных языков, граф зависимостей.
11	Разные подходы к определению понятия смысла языковых единиц: исчисление предикатов.
12	Разные подходы к определению понятия смысла языковых единиц: дистрибутивная семантика.
13	Кластеризация.
14	Нейросетевые подходы к векторному представлению слов (word embedding).
15	Разрешение кореферентности.
16	Понятие трансформера, энкодера, декодера, механизм внимания, современные нейросетевые модели.
17	Машинный перевод, генерация текста.

6.3 График текущего контроля успеваемости

Неделя	Темы занятий	Вид контроля
1	Задача классификации	
2		
3		
4		Коллоквиум
5	Работа с последовательностями	
6		
7		Коллоквиум
8	Теория формальных языков в приложении к естественным	
9		Коллоквиум
10	Семантика	
11		
12		
13		Коллоквиум
14	Трансформеры последовательностей	
15		
16		Коллоквиум

6.4 Методика текущего контроля

на лекционных занятиях текущий контроль включает в себя:

- контроль посещаемости (не более 20% (баллов) от общего объема оценивания текущей аттестации);
- контроль активности студентов. В ходе проведения занятий происходит привлечение студентов к активному участию в дискуссиях, решении задач, обсуждениях и т. д. При этом активность студентов учитывается преподавателем, как один из параметров текущего контроля на практических занятиях (не более 5% (баллов) от общего объема оценивания текущей аттестации).

на практических занятиях текущий контроль включает в себя:

- контроль посещаемости (не более 20% (баллов) от общего объема оценивания текущей аттестации);
- контроль активности студентов. В ходе проведения практических занятий

происходит привлечение студентов к активному участию в дискуссиях, решении задач, обсуждениях и т. д. При этом активность студентов учитывается преподавателем, как один из параметров текущего контроля на практических занятиях (не более 5% (баллов) от общего объема оценивания текущей аттестации);

- распределенный коллоквиум - 5 коллоквиумов по тематике дисциплины (не более 50% (баллов) от общего объема оценивания текущей аттестации). Для допуска к экзамену студенту необходимо решить задачи, выданные в течение семестра. Из каждой темы курса должна быть решена хотя бы одна задача. Список задач формируется на основе актуальной проблематики, связанной с обработкой естественных языков, и обновляется каждый семестр.

По результатам текущего контроля (выполнения всех параметров **более чем на 60 %** (баллов)) студент получает допуск на экзамен.

Контроль самостоятельной работы студентов

Контроль самостоятельной работы студентов осуществляется на лекционных и практических занятиях студентов по методикам, описанным выше.

7 Описание информационных технологий и материально-технической базы

Тип занятий	Тип помещения	Требования к помещению	Требования к программному обеспечению
Лекция	Лекционная аудитория	Количество посадочных мест – в соответствии с контингентом, рабочее место преподавателя, меловая или маркерная доска	
Практические занятия	Аудитория	Количество посадочных мест – в соответствии с контингентом, рабочее место преподавателя, меловая или маркерная доска	

Самостоятельная работа	Помещение для самостоятельной работа	Оснащено компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду университета.	1) Windows 7 и выше или дистрибутив Linux, основанный на Ubuntu или Fedora; 2) Microsoft Office 2007 и выше или Libre Office 6.0 и выше.

8 Адаптация рабочей программы для лиц с ОВЗ

Адаптированная программа разрабатывается при наличии заявления со стороны обучающегося (родителей, законных представителей) и медицинских показаний (рекомендациями психолого-медико-педагогической комиссии). Для инвалидов адаптированная образовательная программа разрабатывается в соответствии с индивидуальной программой реабилитации.