

МИНОБРНАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Институт прикладной математики и компьютерных наук

УТВЕРЖДАЮ
Директор института прикладной
математики и компьютерных наук
А.В. Замятин
« 11 » ноября 2021 г.



Обработка естественного языка

рабочая программа дисциплины

Закреплена за кафедрой Учебный план	<i>Теоретических основ информатики 02.03.02 Фундаментальная информатика и информационные технологии</i>
	<i>Направленность (профиль) «Искусственный интеллект и разработка программных продуктов»</i>
Форма обучения	<i>очная</i>
Общая трудоёмкость	<i>3 з.е.</i>
Часов по учебному плану	<i>108</i>
в том числе:	
аудиторная контактная работа	<i>50.4</i>
самостоятельная работа	<i>57.6</i>
Вид контроля в семестрах зачет	<i>7 семестр – зачет</i>

Программу составил:
канд. техн. наук, доцент,
доцент кафедры теоретических основ информатики

М.С. Пожидаев

Рецензент:
д-р. техн. наук., профессор
профессор кафедры теоретических основ информатики

Ю.Л. Костюк

Рабочая программа дисциплины «Обработка естественного языка» разработана в соответствии с образовательным стандартом высшего образования – бакалавриат, самостоятельно устанавливаемым федеральным государственным автономным образовательным учреждением высшего образования «Национальный исследовательский Томский государственный университет» по направлению подготовки 02.03.02 Фундаментальная информатика и информационные технологии (Утвержден Ученым советом НИ ТГУ, протокол от 27.10.2021 г. № 08).

Рабочая программа одобрена на заседании кафедры теоретических основ информатики

Протокол от 04 июня 2021 г. № 05

Заведующий кафедрой теоретических основ информатики,
д-р техн. наук, профессор

А.В. Замятин

Рабочая программа одобрена на заседании учебно-методической комиссии института прикладной математики и компьютерных наук (УМК ИПМКН)

Протокол от 17 июня 2021 г. № 05

Председатель УМК ИПМКН,
д-р техн. наук, профессор

С.П. Сущенко

Цель освоения дисциплины

Цель – обучить студентов передовым методам, моделям, средствам и технологиям компьютерной обработки текстов на естественных языках, дать умение представлять в алгоритмическом виде процессы анализа и синтеза текста.

1. Место дисциплины в структуре ОПОП

Дисциплина «Обработка естественного языка» относится к части, формируемой участниками образовательных отношений Блока 1 «Дисциплины», входит в модуль «Искусственный интеллект».

Пререквизиты дисциплины: «Основы программирования», «Алгоритмы и структуры данных», «Интеллектуальные системы», «Визуализация многомерных данных», «Статистические методы машинного обучения»

Постреквизиты дисциплины: «Прикладные аспекты машинного обучения», «Преддипломная практика»

2. Компетенции и результаты обучения, формируемые в результате освоения дисциплины

Таблица 1.

Компетенция	Индикатор универсальной компетенции	Код и наименование результатов обучения (планируемые результаты обучения по дисциплине, характеризующие этапы формирования компетенций)
ОПК-2. Способен применять компьютерные/супер компьютерные методы, современное программное обеспечение, в том числе отечественного происхождения, для решения задач профессиональной деятельности	ИОПК-2.1. Обладает необходимыми знаниями основных концепций современных вычислительных систем	ОР-2.1.1. Знать процедуры выявления, формирования и согласования требований к результатам аналитических работ с применением технологий обработки естественного языка;
	ИОПК-2.2. Использует методы высокопроизводительных вычислительных технологий, современного программного обеспечения, в том числе отечественного происхождения	ОР-2.2.1. Знать принципы планирования и организации аналитических работ с использованием технологий обработки естественного языка;
	ИОПК-2.3. Использует инструментальные средства высокопроизводительных вычислений в научной и практической деятельности	ОР-2.3.1. Уметь подготавливать данные для проведения аналитических работ по исследованию текстовых массивов данных методами обработки естественного языка;

ПК-1. Способен осуществлять программирование, тестирование и опытную эксплуатацию ИС с использованием технологических и функциональных стандартов, современных моделей и методов оценки качества и надежности программных средств	ИПК-1.3. Кодирует на языках программирования и проводит модульное тестирование ИС	ОР-1.3.1. Уметь проводить аналитическое исследование и разрабатывать приложения для решения задач обработки естественного языка соответствии с требованиями заказчика
---	---	---

3. Структура и содержание дисциплины

3.1. Структура и трудоемкость видов учебной работы по дисциплине

Общая трудоемкость дисциплины составляет 3 зачетных единицы, 108 часов.

Таблица 2.

Вид учебной работы	Трудоемкость в академических часах	
Общая трудоемкость	108	108
Контактная работа:	50.4	50.4
Лекции (Л):	16	16
Практики (ПЗ)	32	32
Групповые консультации	2.15	2.15
Промежуточная аттестация	0.25	0.25
Самостоятельная работа обучающегося:	57.6	57.6
<i>- изучение учебного материала, публикаций</i>	20	20
<i>- подготовка к практическим занятиям</i>	20	20
<i>- подготовка к рубежному контролю по теме/разделу</i>	17.6	17.6
Вид промежуточной аттестации (зачет, зачет с оценкой, экзамен)	Зачет	Зачет

3.2. Содержание и трудоемкость разделов дисциплины

Таблица 3.

Код занятия	Наименование разделов и тем и их содержание	Вид учебной работы, занятий, контроля	Семестр	Часы в электронной форме	Всего (час.)	Литература	Код (ы) результата(ов) обучения
1	Раздел 1. Введение, история развития дисциплины, решаемые задачи, подходы, методы и инструменты	Л, ПЗ	7		6	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	ОР-2.1.1, ОР-2.2.1, ОР-2.3.1, ОР-1.3.1
		СРС			5		
2	Раздел 2. Предварительна обработка текстовых данных	Л, ПЗ	7		6	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	ОР-2.1.1, ОР-2.2.1, ОР-2.3.1, ОР-1.3.1
		СРС			5		
3	Раздел 3. Разметка последовательностей	Л, ПЗ	7		6	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	ОР-2.1.1, ОР-2.2.1, ОР-2.3.1, ОР-1.3.1
		СРС			5		
4	Раздел 4. Векторизация текстовых данных	Л, ПЗ	7		6	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	ОР-2.1.1, ОР-2.2.1, ОР-2.3.1, ОР-1.3.1
		СРС			5		
5	Раздел 5. Тематическое моделирование	Л, ПЗ	7		6	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	ОР-2.1.1, ОР-2.2.1, ОР-2.3.1, ОР-1.3.1
		СРС			5		
6	Раздел 6. Языковые модели	Л, ПЗ	7		6	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	ОР-2.1.1, ОР-2.2.1, ОР-2.3.1, ОР-1.3.1
		СРС			5		
7	Раздел 7. Анализ тональности	Л, ПЗ	7		6	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	ОР-2.1.1, ОР-2.2.1, ОР-2.3.1, ОР-1.3.1
		СРС			5		
8	Раздел 8. Модели sequence-to-sequence и механизм внимания	Л, ПЗ	7		6	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	ОР-2.1.1, ОР-2.2.1, ОР-2.3.1, ОР-1.3.1
		СРС			5		
	Консультации в период теоретического обучения	Консультации	7		2.15		
	Подготовка к промежуточной аттестации в форме зачета	СРС	7		17.6	x	
	Прохождение промежуточной аттестации в форме зачета	3	7		0.25		

4. Образовательные технологии, учебно-методическое и информационное обеспечение для освоения дисциплины

Занятия по дисциплине проводятся в классической форме в виде лекций и практических занятий:

- лекция – в виде систематического и последовательного изложения преподавателем основного теоретического материала. В начале занятия даётся аннотация излагаемого раздела(лекции). В конце лекции подводится краткий итог изложенного на лекции материала.

- практическое занятие – под руководством преподавателя студенты с использованием предоставленных программно-технических средств и исходных данных выполняют специальные задания, направленные на углубление и закрепление полученных по предмету знаний; работа состоит из экспериментально-практической части и интерпретации полученных результатов на основе теоретических знаний.

Типовые контрольные задания или иные материалы, необходимые для оценки результатов обучения, характеризующих этапы формирования компетенций, и методические материалы, определяющие процедуры оценивания результатов обучения, приведены в Приложении 1 к рабочей программе «Фонд оценочных средств».

4.1. Рекомендуемая литература и учебно-методическое обеспечение

№ п/п	Авторы / составители	Заглавие	Издательство	Год издания, количество страниц
Основная литература				
1.	Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В.	Автоматическая обработка текстов на естественном языке и компьютерная лингвистика	МИЭМ	2011 г., 272 с.
2.	Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С.	Автоматическая обработка текстов на естественном языке и анализ данных	НИУ ВШЭ	2017 г., 269 с.
3		Введение в когнитивную лингвистику	БФУ им. И. Канта	2012 г., 313 с
4	Николенко С., Кадурич А., Архангельская Е.	Глубокое обучение	Питер	2018 г., 480 с
5	Хобсон Лейн, Ханнес Хапке, Коул Ховард	Обработка естественного языка в действии	Питер	2020. г., 576 с
6	Li Deng Yang Liu	Deep Learning in Natural Language Processing		
7	Николаев И.С., Митренина О.В., Ландо Т.М.	. Прикладная и КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА	URSS	2017 г., 320 с
8	Ян Гудфеллоу, Йошуа Бенджио, Аарон Курвилль	Глубокое обучение. Второе цветное издание, исправленное	ДМК Пресс	2018 г., 652 с.
9	Франсуа Шолле	. Глубокое обучение на Python	Питер	2018 г., 400 с
10	Daniel Jurafsky, James H. Martin	Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition	Prentice Hall,	2008 г., 1044 с

4.2. Базы данных и информационно-справочные системы, в том числе зарубежные

4.3. Перечень лицензионного и программного обеспечения

Microsoft Visual Studio или другие среды разработки

4.4. Оборудование и технические средства обучения

Аудитории для проведения интерактивных лекций: видеопроектор, экран настенный.

Для проведения лабораторных занятий – компьютерные классы.

5. Методические указания обучающимся по освоению дисциплины

6. Преподавательский состав, реализующий дисциплину

Пожидаев Михаил Сергеевич, канд. техн. наук, доцент кафедры теоретических основ информатики

7. Язык преподавания – русский язык.