СТЕРЛИТАМАКСКИЙ ФИЛИАЛ
ФЕДЕРАЛЬНОГО ГОСУДАРСТВЕННОГО БЮДЖЕТНОГО ОБРАЗОВАТЕЛЬНОГО
УЧРЕЖДЕНИЯ ВЫСШЕГО ОБРАЗОВАНИЯ
«УФИМСКИЙ УНИВЕРСИТЕТ НАУКИ И ТЕХНОЛОГИЙ»

Факультет ***Математики и информационных технологий***
Кафедра ***Математического моделирования***

Рабочая программа дисциплины (модуля)

дисциплина ***Б1.В.10 Основы обработки текстов***

часть, формируемая участниками образовательных отношений

Направление

01.03.02 ***Прикладная математика и информатика***
код наименование направления

Программа

Искусственный интеллект и анализ данных

Форма обучения

Очная

Для поступивших на обучение в
2023 г.

Разработчик (составитель)
доцент, кандидат физико-математических наук, доцент
Акимов А. А.
ученая степень, должность, ФИО

Стерлитамак 2023

1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с установленными в образовательной программе индикаторами достижения компетенций

Формируемая компетенция (с указанием кода)	Код и наименование индикатора достижения компетенции	Результаты обучения по дисциплине (модулю)
ПК-9. Способен создавать и внедрять одну или несколько сквозных цифровых субтехнологий искусственного интеллекта	ПК-9.1. Участвует в реализации проектов в области сквозной цифровой субтехнологии «Компьютерное зрение»	Обучающийся должен обладать навыками в области машинного обучения, глубокого обучения, обработки изображений, а также понимать принципы работы алгоритмов компьютерного зрения. Он будет применять соответствующие методы и техники, чтобы создать системы, которые могут распознавать и анализировать визуальные данные
	ПК-9.2. Участвует в реализации проектов в области сквозной цифровой субтехнологии «Обработка естественного языка»	Обучающийся должен быть знаком с различными методами и алгоритмами, используемыми в обработке естественного языка, такими как статистический анализ текстов, морфологический анализ, синтаксический анализ, семантический анализ и машинное обучение
	ПК-9.3. Участвует в реализации проектов в различных областях сквозной цифровой субтехнологии	Обучающий должен быть знаком с основами различных субтехнологий искусственного интеллекта, таких как машинное обучение, компьютерное зрение, обработка естественного языка и другие

2. Цели и место дисциплины (модуля) в структуре образовательной программы

Цели изучения дисциплины:
Цели изучения дисциплины фундаментальная подготовка в области основ обработки текстов, овладение средствами обработки больших данных.
Дисциплина относится к базовой и обязательной части изучения.

Дисциплина изучается на 4 курсе в 7 семестре

3. Объем дисциплины (модуля) в зачетных единицах с указанием количества академических или астрономических часов, выделенных на контактную работу обучающихся с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающихся

Общая трудоемкость (объем) дисциплины составляет 4 зач. ед., 144 акад. ч.

Объем дисциплины	Всего часов
	Очная форма обучения
Общая трудоемкость дисциплины	144

Учебных часов на контактную работу с преподавателем:	
лекций	16
практических (семинарских)	16
лабораторных	16
другие формы контактной работы (ФКР)	1,2
Учебных часов на контроль (включая часы подготовки):	34,8
экзамен	
Учебных часов на самостоятельную работу обучающихся (СР)	60

Формы контроля	Семестры
экзамен	7

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины и трудоемкость по видам учебных занятий (в академических часах)

№ п/п	Наименование раздела / темы дисциплины	Виды учебных занятий, включая самостоятельную работу обучающихся и трудоемкость (в часах)			
		Контактная работа с преподавателем			СР
		Лек	Пр/Сем	Лаб	
1	**Распознавание именованных сущностей, машинное обучение с преподавателем**	**2**	**2**	**2**	**8**
1.1	Основы машинного обучения с преподавателем, линейные классификаторы: метод опорных векторов, логистическая регрессия	1	1	1	4
1.2	Задача распознавания именованных сущностей: постановка задачи, области применения, простейшие алгоритмы решения задачи, методы оценки качества	1	1	1	4
2	**Разметка последовательностей, нейронные сети**	**6**	**6**	**6**	**24**
2.1	Понятие разметки последовательность и на примере задачи распознавания именованных сущностей	1	1	1	4
2.2	Условные случайные поля, алгоритм Витерби	1	1	1	4
2.3	Нейронные сети	1	1	1	4
2.4	Нелинейность и дифференцируемость функций в нейронных сетях	1	1	1	4
2.5	Алгоритм обратного распространения ошибок	1	1	1	4
2.6	Типовые практики обучения нейронных сетей	1	1	1	4
3	**Синонимия: дистрибутивные векторные представления слов**	**3**	**3**	**3**	**12**

3.1	Векторные представления слов. Дистрибутивная гипотеза	1	1	1	4
3.2	Локальные модели векторов слов: continousskip -gram, continousbagofword s. Иерархический softmax и negativesampling	1	1	1	4
3.3	Глобальные модели векторов слов: GloVe	1	1	1	4
4	**Символьные представления слов**	**2**	**2**	**2**	**7**
4.1	Использование символьного состава слов в методах обработки текстов на примере задачи распознавания именованных сущностей	1	1	1	4
4.2	Представления слов на основе мешка символьных последовательностей, попарного кодирования байтов, рекуррентных и сверточных нейронных сетей	1	1	1	3
5	**Базовые задачи обработки текстов.**	**3**	**3**	**3**	**9**
5.1	Сегментация текста: задачи токенизации и определения границ предложений. Задача определения языка текста. Профили языка. NaïveBayes классификатор	1	1	1	3
5.2	Оценка качества классификации. Задача определения частей речи и морфологического анализа. Multi - label классификация	1	1	1	3
5.3	Лемматизация: RippleDownRules и LemmaGen. Грамматическая омонимия. Лемматизация как задача классификации	1	1	1	3
	Итого	**16**	**16**	**16**	**60**

4.2. Содержание дисциплины, структурированное по разделам (темам)

Курс лекционных занятий

№	Наименование раздела / темы дисциплины	Содержание
1	**Распознавание именованных сущностей, машинное обучение с преподавателем**	
1.1	Основы машинного обучения с преподавателем, линейные классификаторы: метод опорных векторов, логистическая регрессия	Введение в машинное обучение с преподавателем Линейные классификаторы Метод опорных векторов (Support Vector Machines, SVM). Логистическая регрессия Оценка и интерпретация моделей
1.2	Задача распознавания именованных сущностей: постановка задачи, области применения, простейшие алгоритмы решения задачи, методы оценки качества	Постановка задачи Области применения Простейшие алгоритмы решения задачи Методы оценки качества
2	**Разметка последовательностей, нейронные сети**	
2.1	Понятие разметки последовательность и на примере задачи распознавания именованных сущностей	Задача распознавания именованных сущностей (Named Entity Recognition, NER) Методы и подходы к решению

		задачи NER Обучение моделей NER Примеры применения NER в реальных задачах
2.2	Условные случайные поля, алгоритм Витерби	Метод максимального правдоподобия для обучения CRF Регуляризация в условных случайных полях Алгоритм прямого-обратного прохода (Forward-Backward Algorithm) в CRF Алгоритм Витерби для нахождения наиболее вероятной последовательности состояний
2.3	Нейронные сети	Искусственные нейроны и их активационные функции Архитектура прямого распространения нейронных сетей Обратное распространение ошибки Оптимизация весов нейронных сетей Регуляризация и предотвращение переобучения Рекуррентные нейронные сети для анализа последовательностей
2.4	Нелинейность и дифференцируемость функций в нейронных сетях	Линейные и нелинейные функции в нейронных сетях Значение нелинейности в активационных функциях Дифференцируемость и ее роль в обучении нейронных сетей Свойства нелинейных функций и их влияние на выразительность нейронных сетей Проблема затухающего градиента и роль нелинейных функций в ее решении Нелинейные функции и сложность обучения нейронных сетей
2.5	Алгоритм обратного распространения ошибок	Обзор алгоритма обратного распространения ошибок Математические основы алгоритма: прямой проход и обратный проход Функции активации и их роли в алгоритме Регуляризация и предотвращение переобучения
2.6	Типовые практики обучения нейронных сетей	Архитектуры нейронных сетей: полносвязные, сверточные, рекуррентные и т.д. Инициализация весов нейронных сетей Функции активации и их роль в

		обучении Процесс обратного распространения ошибки
3	**Синонимия: дистрибутивные векторные представления слов**	
3.1	Векторные представления слов. Дистрибутивная гипотеза	Дистрибутивная гипотеза: основные принципы и идеи Матричные представления слов Word2Vec: Continuous Bag of Words (CBOW) и Skip-gram модели GloVe: Global Vectors for Word Representation FastText: представления слов на основе n-грамм Применение векторных представлений слов в задачах обработки естественного языка
3.2	Локальные модели векторов слов: continousskip -gram, continousbagofword s. Иерархический softmax и negativesampling	Continuous Skip-gram (непрерывный пропуск грамматики) Continuous Bag of Words (непрерывная мешковина слов) Иерархический softmax Negative Sampling (отрицательное сэмплирование)
3.3	Глобальные модели векторов слов: GloVe	Проблема распределенного представления слов Основные принципы и идеи модели GloVe Математические основы GloVe Оценка и интерпретация результатов GloVe Преимущества и ограничения использования GloVe
4	**Символьные представления слов**	
4.1	Использование символьного состава слов в методах обработки текстов на примере задачи распознавания именованных сущностей	Основные методы и подходы к NER Значение символьного состава слов в обработке текстов Преобразование слов в символьные последовательности: токенизация и разбиение на символы Использование символьных эмбеддингов для представления слов Морфологический анализ символьных последовательностей Архитектуры символьных моделей для NER
4.2	Представления слов на основе мешка символьных последовательностей, попарного кодирования байтов, рекуррентных и сверточных нейронных сетей	Мешок символьных последовательностей: концепция и применение Попарное кодирование байтов: принцип работы и особенности Рекуррентные нейронные сети (RNN) для представления слов:

		архитектура и обучение Сверточные нейронные сети (CNN) для представления слов: основные идеи и примеры
5	**Базовые задачи обработки текстов.**	
5.1	Сегментация текста: задачи токенизации и определения границ предложений. Задача определения языка текста. Профили языка. NaïveBayes классификатор	Задача токенизации: определение единиц текста Методы токенизации: разделение на слова, символы, морфемы и т.д. Алгоритмы определения границ предложений Обзор методов и подходов к определению языка текста
5.2	Оценка качества классификации. Задача определения частей речи и морфологического анализа. Multi - label классификация	Метрики оценки качества классификации Матрица ошибок (Confusion Matrix) Точность (Accuracy) Полнота (Recall) Точность и полнота в многоклассовой классификации F-мера (F-measure) ROC-кривая и AUC-ROC
5.3	Лемматизация: RippleDownRules и LemmaGen. Грамматическая омонимия. Лемматизация как задача классификации	Метрики оценки качества классификации: точность, полнота, F-мера, ROC-кривая, площадь под кривой Матрица ошибок и её интерпретация Кросс-валидация и её роль в оценке качества классификации Подбор порогового значения для классификаторов Лемматизация: понятие и задачи

Курс практических/семинарских занятий

№	Наименование раздела / темы дисциплины	Содержание
1	**Распознавание именованных сущностей, машинное обучение с преподавателем**	
1.1	Основы машинного обучения с преподавателем, линейные классификаторы: метод опорных векторов, логистическая регрессия	Обучение и оценка моделей линейной классификации
1.2	Задача распознавания именованных сущностей: постановка задачи, области применения, простейшие алгоритмы решения задачи, методы оценки качества	Применение алгоритмов машинного обучения, таких как условные случайные поля (Conditional Random Fields, CRF), рекуррентные нейронные сети (Recurrent Neural Networks, RNN), сверточные нейронные сети (Convolutional Neural Networks, CNN) и другие, для обучения моделей распознавания именованных сущностей.
2	**Разметка последовательностей, нейронные сети**	
2.1	Понятие разметки последовательность и на примере	Практическое решение задачи NER с использованием выбранной модели и набора

	задачи распознавания именованных сущностей	данных
2.2	Условные случайные поля, алгоритм Витерби	Реализация алгоритма Витерби на практике с использованием различных программных инструментов или языков программирования
2.3	Нейронные сети	Обучение однослойных нейронных сетей методом градиентного спуска
2.4	Нелинейность и дифференцируемость функций в нейронных сетях	Практическое применение различных функций активации в нейронных сетях с использованием фреймворков, таких как TensorFlow или PyTorch
2.5	Алгоритм обратного распространения ошибок	Разработка и реализация нейронной сети с использованием Python и библиотек машинного обучения, например, TensorFlow или PyTorch
2.6	Типовые практики обучения нейронных сетей	Создание и настройка архитектуры нейронной сети
3	**Синонимия: дистрибутивные векторные представления слов**	
3.1	Векторные представления слов. Дистрибутивная гипотеза	Работа с библиотеками и инструментами для создания векторных представлений слов, такими как Gensim, TensorFlow или PyTorch
3.2	Локальные модели векторов слов: continousskip -gram, continousbagofword s. Иерархический softmax и negativesampling	Обучение модели на больших корпусах текста
3.3	Глобальные модели векторов слов: GloVe	Обучение модели GloVe на предварительно подготовленных текстовых данных
4	**Символьные представления слов**	
4.1	Использование символьного состава слов в методах обработки текстов на примере задачи распознавания именованных сущностей	Предобработка текстовых данных: токенизация, нормализация, удаление стоп-слов и другие шаги подготовки данных
4.2	Представления слов на основе мешка символьных последовательностей, попарного кодирования байтов, рекуррентных и сверточных нейронных сетей	Реализация модели на основе мешка символов с использованием библиотеки или фреймворка для глубокого обучения
5	**Базовые задачи обработки текстов.**	
5.1	Сегментация текста: задачи токенизации и определения границ предложений. Задача определения языка текста. Профили языка. NaïveBayes классификатор	Применение алгоритмов машинного обучения для определения границ предложений
5.2	Оценка качества классификации. Задача определения частей речи и морфологического анализа. Multi - label классификация	Реализация и применение различных метрик оценки качества классификации на реальных наборах данных
5.3	Лемматизация: RippleDownRules и LemmaGen. Грамматическая омонимия. Лемматизация как задача классификации	Разработка и реализация алгоритма RippleDownRules для лемматизации

Курс лабораторных занятий

№	Наименование раздела / темы дисциплины	Содержание
1	**Распознавание именованных сущностей, машинное обучение с преподавателем**	
1.1	Основы машинного обучения с преподавателем, линейные классификаторы: метод опорных векторов, логистическая регрессия	Применение стратегий для справления с несбалансированными данными: upsampling, downsampling, взвешивание классов
1.2	Задача распознавания именованных сущностей: постановка задачи, области применения, простейшие алгоритмы решения задачи, методы оценки качества	Кросс-валидация и разделение данных для оценки качества модели
2	**Разметка последовательностей, нейронные сети**	
2.1	Понятие разметки последовательность и на примере задачи распознавания именованных сущностей	Разработка и применение правил и эвристик для улучшения результатов NER
2.2	Условные случайные поля, алгоритм Витерби	Демонстрация применения CRF и алгоритма Витерби на различных сценариях и задач
2.3	Нейронные сети	Генеративные модели
2.4	Нелинейность и дифференцируемость функций в нейронных сетях	Исследование влияния выбора функции активации на скорость сходимости и производительность обучения нейронной сети
2.5	Алгоритм обратного распространения ошибок	Реализация алгоритма обратного распространения ошибок в коде нейронной сети и обучение модели на обучающей выборке
2.6	Типовые практики обучения нейронных сетей	Обучение нейронной сети
3	**Синонимия: дистрибутивные векторные представления слов**	
3.1	Векторные представления слов. Дистрибутивная гипотеза	Применение векторных представлений слов для различных задач обработки естественного языка
3.2	Локальные модели векторов слов: continousskip -gram, continousbagofword s. Иерархический softmax и negativesampling	Реализация и обучение моделей continous skip-gram и continous bag of words на выбранном корпусе текста
3.3	Глобальные модели векторов слов: GloVe	Визуализация пространства векторов слов с помощью методов снижения размерности, таких как t-SNE или PCA
4	**Символьные представления слов**	
4.1	Использование символьного состава слов в методах обработки текстов на примере задачи распознавания именованных сущностей	Обучение и применение моделей машинного обучения для задачи NER, используя символьные признаки. Примеры моделей могут включать рекуррентные нейронные сети (RNN), сверточные нейронные сети (CNN), комбинированные модели и т.д.
4.2	Представления слов на основе мешка символьных	Применение модели на основе мешка символов для задач классификации или

	последовательностей, попарного кодирования байтов, рекуррентных и сверточных нейронных сетей	генерации текста
5	**Базовые задачи обработки текстов.**	
5.1	Сегментация текста: задачи токенизации и определения границ предложений. Задача определения языка текста. Профили языка. NaïveBayes классификатор	Применение классификаторов и моделей машинного обучения для определения языка текста
5.2	Оценка качества классификации. Задача определения частей речи и морфологического анализа. Multi - label классификация	Реализация и применение алгоритмов и моделей для решения задачи multi-label классификации на реальных наборах данных
5.3	Лемматизация: RippleDownRules и LemmaGen. Грамматическая омонимия. Лемматизация как задача классификации	Изучение и применение метода LemmaGen в контексте лемматизации

5. Учебно-методическое обеспечение для самостоятельной работы обучающихся по дисциплине (модулю)

Самостоятельная работа студентов, предусмотренная учебным планом, должна соответствовать более глубокому усвоению изучаемого материала, формировать навыки исследовательской работы и ориентировать их на умение применять полученные теоретические знания на практике. В процессе этой деятельности решаются задачи:
– научить студентов работать с учебной литературой;
– формировать у них соответствующие знания, умения и навыки;
– стимулировать профессиональный рост студентов, воспитывать творческую активность и инициативу.
Самостоятельная работа студентов предполагает:
– подготовку к занятиям (изучение лекционного материала и чтение литературы);
– оформление отчета по самостоятельной работе;
– подготовку к итоговому контролю.

6. Учебно-методическое и информационное обеспечение дисциплины (модуля)

6.1. Перечень учебной литературы, необходимой для освоения дисциплины (модуля)

Основная учебная литература:

1. Михеева Е.В., Информационные технологии в профессиональной деятельности учебное пособие для студентов учреждений СПО/ Е.В. Михеева 11- е изд., стер.- М.: Академия, 2015.- 384 с. (25.06.2023).

Дополнительная учебная литература:

1. Интеллектуальные информационные системы и технологии : учебное пособие / Ю.Ю. Громов, О.Г. Иванова, В.В. Алексеев и др. ; Министерство образования и науки Российской Федерации, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Тамбовский государственный технический университет». - Тамбов : Издательство ФГБОУ ВПО «ТГТУ», 2013. - 244
с. : ил. - Библиогр. в кн. - ISBN 978-5-8265-1178-7; [Электронный ресурс]. - URL:

http://biblioclub.ru/index.php?page=book&id=277713.— Загл. с экрана. (25.06.2023).

6.2. Перечень электронных библиотечных систем, современных профессиональных баз данных и информационных справочных систем

№ п/п	Наименование документа с указанием реквизитов
1	Договор на доступ к ЭБС ZNANIUM.COM между БашГУ в лице директора СФ БашГУ и ООО «Знаниум»№ 3/22-эбс от 05.07.2022
2	Договор на доступ к ЭБС «ЭБС ЮРАЙТ» (полная коллекция) между БашГУ в лице директора СФ БашГУ и ООО «Электронное издательство ЮРАЙТ» № 1/22-эбс от 04.03.2022
3	Договор на доступ к ЭБС «Университетская библиотека онлайн» между БашГУ и «Нексмедиа» № 223-950 от 05.09.2022
4	Договор на доступ к ЭБС «Лань» между БашГУ и издательством «Лань» № 223-948 от 05.09.2022
5	Договор на доступ к ЭБС «Лань» между БашГУ и издательством «Лань» № 223-949 от 05.09.2022
6	Соглашение о сотрудничестве между БашГу и издательством «Лань» № 5 от 05.09.2022
7	ЭБС «ЭБ БашГУ», бессрочный договор между БашГУ и ООО «Открытые библиотечные системы» № 095 от 01.09.2014 г.
8	Договор на БД диссертаций между БашГУ и РГБ № 223-796 от 27.07.2022
9	Договор о подключении к НЭБ и о предоставлении доступа к объектам НЭБ между БашГУ в лице директора СФ БашГУ с ФГБУ «РГБ» № 101/НЭБ/1438-П от 11.06.2019
10	Договор на доступ к ЭБС «ЭБС ЮРАЙТ» (полная коллекция) между УУНиТ в лице директора СФ УУНиТ и ООО «Электронное издательство ЮРАЙТ» № 1/23-эбс от 03.03.2023

Перечень ресурсов информационно-телекоммуникационной сети «Интернет» (далее - сеть «Интернет»)

№ п/п	Адрес (URL)	Описание страницы
1	https://intuit.ru/	Бесплатное дистанционное обучение в национальном открытом институте "Интуит".

6.3. Перечень лицензионного и свободно распространяемого программного обеспечения, в том числе отечественного производства

Наименование программного обеспечения
Office Standart 2007 Russian OpenLicensePack NoLevel Acdmc 200 / ООО «Общество информационных технологий». Государственный контракт №13 от 06.05.2009
Kaspersky Endpoint Security950 /СофтЛайн Трейд, АО №лиц.17E0-171109-063136-757-608
Windows XP Лицензионное соглашение MSDN. Государственный контракт №9 от 18.03.2008 г. ЗАО «СофтЛайн»

7. Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине (модулю)

Тип учебной аудитории	Оснащенность учебной аудитории

Учебная аудитория для проведения занятий лекционного типа, учебная аудитория для проведения занятий семинарского типа, учебная аудитория текущего контроля и промежуточной аттестации, учебная аудитория групповых и индивидуальных консультаций	Доска, учебная мебель, переносной проектор, переносной экран, учебно-наглядные пособия
Учебная аудитория для проведения занятий семинарского типа, учебная аудитория текущего контроля и промежуточной аттестации, учебная аудитория групповых и индивидуальных консультаций	Доска, учебная мебель
Учебная аудитория для проведения занятий лекционного типа, учебная аудитория для проведения занятий семинарского типа, учебная аудитория текущего контроля и промежуточной аттестации, учебная аудитория групповых и индивидуальных консультаций	Доска, учебная мебель, проектор, экран, учебно-наглядные пособия