Министерство науки и высшего образования Российской Федерации

ФГБОУ ВО «БАЙКАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

УТВЕРЖДАЮ
Проректор по учебной работе
д.юр.н., доц. Васильева Н.В.

_____

21.06.2024г.

Рабочая программа дисциплины
Б1.У.10. Обработка текстов на естественных языках

Направление подготовки: 09.03.03 Прикладная информатика
Направленность (профиль): Системы искусственного интеллекта
Квалификация выпускника: бакалавр
Форма обучения: очная, заочная

	Очная ФО	Заочная ФО
Курс	4	4
Семестр	41	41
Лекции (час)	14	6
Практические (сем, лаб.) занятия (час)	28	10
Самостоятельная работа, включая подготовку к экзаменам и зачетам (час)	66	92
Курсовая работа (час)		
Всего часов	108	108
Зачет (семестр)		
Экзамен (семестр)	41	41

Иркутск 2024

Программа составлена в соответствии с ФГОС ВО по направлению 09.03.03 Прикладная информатика.

Автор Е.В. Аксенюшкина

Рабочая программа обсуждена и утверждена на заседании кафедры математических методов и цифровых технологий

Заведующий кафедрой А.В. Родионов

1. Цели изучения дисциплины

Цели освоения дисциплины: выработка у студентов компетенций, связанных с их теоретической и практической подготовкой к использованию средств и методов анализа и математического моделирования структур естественного языка.

Задачи:
- овладеть теоретическими знаниями и практическими навыками использования средств и методов анализа и математического моделирования структур естественного языка;
- разъяснить ограничения и особенности применения различных методов анализа и математического моделирования структур естественного языка;
- практическое применение современных программных средств и специализированных библиотек для обработки текстов.

2. Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения образовательной программы

Компетенции обучающегося, формируемые в результате освоения дисциплины

Код компетенции по ФГОС ВО	Компетенция
ПК-8	Способен руководить проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях

Структура компетенции

Компетенция	Формируемые ЗУНы
ПК-8 Способен руководить проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях	З. Знать, как руководить проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях У. Уметь разрабатывать проекты по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях Н. Владеть навыками руководства проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях

3. Место дисциплины (модуля) в структуре образовательной программы

Принадлежность дисциплины - БЛОК 1 ДИСЦИПЛИНЫ (МОДУЛИ): Часть, формируемая участниками образовательных отношений.

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

Общая трудоемкость дисциплины составляет 3 зач. ед.,108 часов.

Вид учебной работы	Количество часов (очная ФО)	Количество часов (заочная ФО)
Контактная(аудиторная) работа		
Лекции	14	6
Практические (сем, лаб.) занятия	28	10
Самостоятельная работа, включая подготовку к экзаменам и зачетам	66	92
Всего часов	108	108

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

5.1. Содержание разделов дисциплины

Заочная форма обучения

№ п/п	Раздел и тема дисциплины	Семестр	Лекции	Семинар Лаборат. Практич.	Самостоят. раб.	В интерактивной форме	Формы текущего контроля успеваемости
1	Введение в обработку текста	41	1	2	20		
2	Предобработка текста	41	2	2	26		Лабораторная работа №1 по темам 1, 2
3	Извлечение и анализ данных	41	2	4	26		Лабораторная работа №2 по теме 3. Лабораторная работа №3 по теме 3
4	Языковые модели	41	1	2	20		Лабораторная работа №4 по темам 3, 4. Теоретический тест по темам 1 - 4
	ИТОГО		6	10	92		

Очная форма обучения

№ п/п	Раздел и тема дисциплины	Семестр	Лекции	Семинар Лаборат. Практич.	Самостоят. раб.	В интерактивной форме	Формы текущего контроля успеваемости
1	Введение в обработку текста	41	2	2	13		
2	Предобработка текста	41	4	4	20		Лабораторная работа №1 по темам 1, 2
3	Извлечение и анализ данных	41	6	16	20		Лабораторная работа №2 по теме 3. Лабораторная работа №3 по теме 3
4	Языковые модели	41	2	6	13		Лабораторная

№ п/п	Раздел и тема дисциплины	Семе-стр	Лек-ции	Семинар Лаборат. Практич.	Само-стоят. раб.	В интера-ктивной форме	Формы текущего контроля успеваемости
							работа №4 по темам 3, 4. Теоретический тест по темам 1 - 4
	ИТОГО		14	28	66		

5.2. Лекционные занятия, их содержание

№ п/п	Наименование разделов и тем	Содержание
1	Введение в NLP	Применение языка Python для обработки естественного языка. Применение машинного обучения для обработки естественного языка. Нейросетевые модели. Использование сверточных сетей для NLP
2	Конвейер обработки текста	Установка статистических моделей в библиотеки spaCy. Токенизация. Лемматизация. Частеречная разметка. Синтаксические отношения. Распознавание именованных сущностей.
3	Работа с объектами-контейнерами и настойка spaCy	Объекты-контейнеры библиотеки spaCy. Настройка обработки текста. Использование структур данных уровня языка C библиотеки spaCy.
4	Выделение и использование лингвистических признаков	Выделение и генерация текста с помощью тегов частей речи. Использование меток синтаксических зависимостей при обработке текста
5	Работа с векторами слов	Смысл векторов слов. Установка пакетов векторов слов. Сравнение объектов spaCy.
6	Поиск паттернов и обход деревьев зависимостей	Паттерны последовательности слов. Создание паттернов на основе пользовательских признаков. Применение паттернов последовательностей слов в чат-ботах для генерации высказываний. Выделение информации путем обхода дерева зависимостей. Каткое изложение текста с помощью дерева зависимостей.
7	Обучение моделей	Обучение компонента конвейера модели. Обучение средства распознавание именованных сущностей. Создание обучающих примеров данных. Автоматизация процесса создания примеров данных. Процесс обучения. Создание нового синтаксического анализатора

5.3. Семинарские, практические, лабораторные занятия, их содержание

№ раздела и темы	Содержание и формы проведения
1	Введение в обработку естественных языков. Обзор основных понятий и задач NLP. Установка необходимых библиотек.
2	Предобработка текста. Токенизация текста. Удаление стоп-слов и пунктуации.
2	Предобработка текста. Преобразование текста: стемминг и лемматизация.

№ раздела и темы	Содержание и формы проведения
	Практика с NLTK и pymorphy2.
3	Векторизация текста. Обзор методов векторизации: Bag of Words, TF-IDF. Практическое применение с использованием sklearn.
3	Введение в машинное обучение для NLP. Обзор моделей машинного обучения, применяемых в NLP.
3	Классификация текстов. Построение и оценка моделей классификации текстов (например, спам/не спам). Работа с наивным байесовским классификатором и SVM.
3	Кластеризация текстовых данных. Методы кластеризации текстов (k-means, иерархическая кластеризация). Визуализация результатов кластеризации.
3	Извлечение признаков из текста. Продвинутые методы извлечения признаков (World2Vec, GloVe). Практика с библиотекой Gensim.
3	Распознавание именованных сущностей (NER). Практическое занятие по использованию моделей NER. Работа с библиотеками Natasha и SpaCy.
3	Синтаксический анализ. Парсинг зависимостей и построение синтаксических деревьев. Практика с библиотекой SpaCy.
3	Анализ настроений. Построение моделей для определения настроения текста. Сравнение различных подходов и анализ результатов.
4	Тематическое моделирование. Введение в LDA и другие методы тематического моделирования.
4	Работа с естественно-языковыми интерфейсами. Создание чат-ботов и вопросно-ответных систем. Интеграция с Telegram или другими мессенджерами.
4	Машинный перевод. Обзор технологий и алгоритмов машинного перевода. Практическое занятие с использованием Google Translate API или других сервисов.

6. Фонд оценочных средств для проведения промежуточной аттестации по дисциплине (полный текст приведен в приложении к рабочей программе)

6.1. Текущий контроль

№ п/п	Этапы формирования компетенций (Тема из рабочей программы дисциплины)	Перечень формируемых компетенций по ФГОС ВО	(ЗУНы: (З.1…З.n, У.1…У.n, Н.1…Н.n)	Контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы (Наименование оценочного средства)	Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания (по 100-балльной шкале)
1	2. Предобработка текста	ПК-8	У.Уметь разрабатывать проекты по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях	Лабораторная работа №1 по темам 1, 2	Лабораторная работа №1 оценивается в 20 баллов (20)

№ п/п	Этапы формирования компетенций (Тема из рабочей программы дисциплины)	Перечень формируемых компетенций по ФГОС ВО	(ЗУНы: (З.1…З.n, У.1…У.n, Н.1…Н.n)	Контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы (Наименование оценочного средства)	Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания (по 100-балльной шкале)
			Н.Владеть навыками руководства проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях		
2	3. Извлечение и анализ данных	ПК-8	У.Уметь разрабатывать проекты по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях Н.Владеть навыками руководства проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях	Лабораторная работа №2 по теме 3	Лабораторная работа №2 оценивается в 20 баллов (20)
3		ПК-8	У.Уметь разрабатывать проекты по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях Н.Владеть навыками руководства проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий	Лабораторная работа №3 по теме 3	Лабораторная работа №3 оценивается в 20 баллов (20)

№ п/п	Этапы формирования компетенций (Тема из рабочей программы дисциплины)	Перечень формируемых компетенций по ФГОС ВО	(ЗУНы: (З.1…З.n, У.1…У.n, Н.1…Н.n)	Контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы (Наименование оценочного средства)	Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания (по 100-балльной шкале)
			искусственного интеллекта в прикладных областях		
4	4. Языковые модели	ПК-8	У.Уметь разрабатывать проекты по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях Н.Владеть навыками руководства проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях	Лабораторная работа №4 по темам 3, 4	Лабораторная работа №4 оценивается в 20 баллов (20)
5		ПК-8	З.Знать, как руководить проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях	Теоретический тест по темам 1 - 4	Теоретический тест состоит из 10 вопрос. Правильный ответ на вопрос оценивается в 2 балла. (20)
				Итого	**100**

6.2. Промежуточный контроль (зачет, экзамен)

Рабочим учебным планом предусмотрен Экзамен в семестре 41.

ВОПРОСЫ ДЛЯ ПРОВЕРКИ ЗНАНИЙ:

1-й вопрос билета (30 баллов), вид вопроса: Тест/проверка знаний. Критерий: Правильный ответ на каждый вопрос оценивается в 3 балла.

Компетенция: ПК-8 Способен руководить проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях

Знание: Знать, как руководить проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях

1. Дистрибутивно-семантические модели.
2. Естественные языки, особенности обработки естественного языка.
3. Метод Continuous Bag of Words.
4. Метод GloVe.
5. Метод s построения векторной модели теста: мешок слов, n-граммы.
6. Метод Skip-Gram.
7. Метод T-IDF.
8. Методы преобразования последовательностей.
9. Методы решения задачи выделения фрагментов текста и их соотнесения с заданными классам.
10. Моделирование языка.
11. Модель FastText
12. Модель Word2Vec.
13. Основные задачи обработки текстов на естественном языке: лингвистический анализ, извлечение признаков из текстов, прикладные задачи обработки текстов, генерация текста.
14. Применение рекуррентных нейронных сетей для решения задачи генерации текстов.
15. Примеры применения сверточных нейронных сетей для решения задач обработки текстов.
16. Программные продукты и библиотеки, используемые для построения векторной модели текста.

ТИПОВЫЕ ЗАДАНИЯ ДЛЯ ПРОВЕРКИ УМЕНИЙ:

2-й вопрос билета (30 баллов), вид вопроса: Задание на умение. Критерий: Правильно выполненное задание оценивается в 30 баллов.

Компетенция: ПК-8 Способен руководить проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях

Умение: Уметь разрабатывать проекты по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях

Задача № 1. Выполните базовую предобработку текста

ТИПОВЫЕ ЗАДАНИЯ ДЛЯ ПРОВЕРКИ НАВЫКОВ:

3-й вопрос билета (40 баллов), вид вопроса: Задание на навыки. Критерий: Правильно выполненное задание оценивается в 40 баллов.

Компетенция: ПК-8 Способен руководить проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях

Навык: Владеть навыками руководства проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых технологий искусственного интеллекта в прикладных областях

Задание № 1. Проведите классификацию текста

ОБРАЗЕЦ БИЛЕТА

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«БАЙКАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «БГУ»)

Направление - 09.03.03 Прикладная информатика
Профиль - Системы искусственного интеллекта
Кафедра математических методов и цифровых технологий
Дисциплина - Обработка текстов на естественных языках

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 1

1. Тест (30 баллов).
2. Выполните базовую предобработку текста (30 баллов).
3. Проведите классификацию текста (40 баллов).

Составитель _____ Е.В. Аксенюшкина

Заведующий кафедрой _____ А.В. Родионов

7. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля)

а) основная литература:

1. Бурдина Н. Г., Новгородцева Т. Ю. Технологии обработки текстовой информации. практикум по Word/ сост. Н. Г. Бурдина, Т. Ю. Новгородцева.- Иркутск: Изд-во БГУЭП, 2006.-84 с.
2. Буйначев, С. К. Основы программирования на языке Python [Электронный ресурс] : учебное пособие / С. К. Буйначев, Н. Ю. Боклаг ; под ред. Ю. В. Песин. — Электрон. текстовые данные. — Екатеринбург : Уральский федеральный университет, ЭБС АСВ, 2014. — 92 с. — 978-5-7996-1198-9. — Режим доступа: http://www.iprbookshop.ru/66183.html
3. Гольдберг, Й. Нейросетевые методы в обработке естественного языка / Й. Гольдберг ; перевод А. А. Слинкин. — Москва : ДМК Пресс, 2019. — 282 с. — ISBN 978-5-97060-754-1. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/124564.html

б) дополнительная литература:

1. Информационные технологии работы с текстами. учеб. пособие/ А. В. Бурдуковская [и др.].- Иркутск: Изд-во БГУЭП, 2003.-61 с.
2. Ганегедара, Т. Обработка естественного языка с TensorFlow / Т. Ганегедара ; перевод В. С. Яценков. — Москва : ДМК Пресс, 2020. — 382 с. — ISBN 978-5-97060-756-5. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/130336.html

8. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины (модуля), включая профессиональные базы данных и информационно-справочные системы

Для освоения дисциплины обучающемуся необходимы следующие ресурсы информационно-телекоммуникационной сети «Интернет»:
– Сайт Байкальского государственного университета, адрес доступа: http://bgu.ru/, доступ круглосуточный неограниченный из любой точки Интернет
– Электронно-библиотечная система IPRbooks, адрес доступа: https://www.iprbookshop.ru. доступ неограниченный

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

Изучать дисциплину рекомендуется в соответствии с той последовательностью, которая обозначена в ее содержании. Для успешного освоения курса обучающиеся должны иметь первоначальные знания в области анализа данных.

На лекциях преподаватель озвучивает тему, знакомит с перечнем литературы по теме, обосновывает место и роль этой темы в данной дисциплине, раскрывает ее практическое значение. В ходе лекций студенту необходимо вести конспект, фиксируя основные понятия и проблемные вопросы.

Практические (семинарские) занятия по своему содержанию связаны с тематикой лекционных занятий. Начинать подготовку к занятию целесообразно с конспекта лекций. Задание на практическое (семинарское) занятие сообщается обучающимся до его проведения. На семинаре преподаватель организует обсуждение этой темы, выступая в качестве организатора, консультанта и эксперта учебно-познавательной деятельности обучающегося.

Изучение дисциплины (модуля) включает самостоятельную работу обучающегося.

Основными видами самостоятельной работы студентов с участием преподавателей являются:

• текущие консультации;

• прием и разбор домашних заданий (в часы практических занятий);

• прием и защита лабораторных работ (во время проведения занятий);.

Основными видами самостоятельной работы студентов без участия преподавателей являются:

• формирование и усвоение содержания конспекта лекций на базе рекомендованной лектором учебной литературы, включая информационные образовательные ресурсы (электронные учебники, электронные библиотеки и др.);

• самостоятельное изучение отдельных тем или вопросов по учебникам или учебным пособиям;

• подготовка к семинарам и лабораторным работам;

• выполнение домашних заданий в виде решения отдельных задач, проведения типовых расчетов, расчетно-компьютерных и индивидуальных работ по отдельным разделам содержания дисциплин и др.

10. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения

В учебном процессе используется следующее программное обеспечение:
– MS Office,
– Python,

11. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю):

В учебном процессе используется следующее оборудование:

– Помещения для самостоятельной работы, оснащенные компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду вуза,
– Учебные аудитории для проведения: занятий лекционного типа, занятий семинарского типа, практических занятий, выполнения курсовых работ, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, укомплектованные специализированной мебелью и техническими средствами обучения,
– Компьютерный класс