Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования

Пермский национальный исследовательский политехнический университет

УТВЕРЖДАЮ
Проректор по образовательной деятельности

_____ А.Б. Петроченков

« 07 » марта 20 23 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Дисциплина: Модели и методы обработки естественного языка
(наименование)

Форма обучения: очная
(очная/очно-заочная/заочная)

Уровень высшего образования: магистратура
(бакалавриат/специалитет/магистратура)

Общая трудоёмкость: 144 (4)
(часы (ЗЕ))

Направление подготовки: 09.04.04 Программная инженерия
(код и наименование направления)

Направленность: Разработка программно-информационных систем
(наименование образовательной программы)

Пермь 2023

1. Общие положения

1.1. Цели и задачи дисциплины

Формирование комплекса знаний, умений и навыков в области применения моделей, методов и инструментов обработки естественного языка для разработки информационного, лингвистического, программного обеспечения интеллектуальных информационных систем.

1.2. Изучаемые объекты дисциплины

Естественный язык; цифровое представление естественного языка; модели естественного языка; методы обработки естественного языка; программные инструменты обработки естественного языка.

1.3. Входные требования

Не предусмотрены

2. Планируемые результаты обучения по дисциплине

Компетенция	Индекс индикатора	Планируемые результаты обучения по дисциплине (знать, уметь, владеть)	Индикатор достижения компетенции, с которым соотнесены планируемые результаты обучения	Средства оценки
ПК-2.1	ИД-1ПК-2.1	Знает методологии разработки программного обеспечения обработки естественного языка	Знает методологии разработки программного обеспечения	Дифференцированный зачет
ПК-2.1	ИД-2ПК-2.1	Умеет применять нормативные документы, определяющие требования к оформлению программного кода информационных систем обработки естественного языка	Умеет применять нормативные документы, определяющие требования к оформлению программного кода	Отчёт по практическому занятию
ПК-2.1	ИД-3ПК-2.1	Владеет навыками оценки качества и эффективности программного кода в информационных системах обработки естественного языка	Владеет навыками оценки качества и эффективности программного кода	Защита лабораторной работы

3. Объем и виды учебной работы

Вид учебной работы	Всего часов	Распределение по семестрам в часах
		Номер семестра
		4
1. Проведение учебных занятий (включая проведение текущего контроля успеваемости) в форме: 1.1. Контактная аудиторная работа, из них:	72	72
- лекции (Л)	18	18
- лабораторные работы (ЛР)	24	24
- практические занятия, семинары и (или) другие виды занятий семинарского типа (ПЗ)	26	26
- контроль самостоятельной работы (КСР)	4	4
- контрольная работа		
1.2. Самостоятельная работа студентов (СРС)	72	72
2. Промежуточная аттестация		
Экзамен		
Дифференцированный зачет	9	9
Зачет		
Курсовой проект (КП)		
Курсовая работа (КР)		
Общая трудоемкость дисциплины	144	144

4. Содержание дисциплины

Наименование разделов дисциплины с кратким содержанием	Объем аудиторных занятий по видам в часах			Объем внеаудиторных занятий по видам в часах
	Л	ЛР	ПЗ	СРС
4-й семестр				
Введение в обработку естественного языка	4	2	0	6
Понятие естественного языка. Характеристики естественного языка. Уровни языка. Этапы анализа текста. Основные задачи обработки естественного языка. Языковые модели. Морфологический анализ и синтез. Стемминг, лемматизация, полный морфоанализ. Морфологические процессоры. Инструменты обработки естественного языка в экосистеме Python.				
Векторные модели текста	2	2	4	8
Методы сбора текстовых данных. Предварительная обработка текста. Извлечение языковых данных. Работа с N-граммами. Лемматизация. Векторизация. Обзор подходов к векторизации. Метрики. Векторная модель документа. Использование векторного представления для анализа текстов. Работа в N-мерном пространстве.				

Наименование разделов дисциплины с кратким содержанием	Объем аудиторных занятий по видам в часах			Объем внеаудиторных занятий по видам в часах
	Л	ЛР	ПЗ	СРС
Информационный поиск	2	4	4	10
Индексирование. Булевый поиск, ранжированный поиск. Оценка релевантности документа. Принципы работы поисковых движков. Квазиреферирование и автоматическое аннотирование документов. Основные стратегии квазиреферирования. Обзорное реферирование. Типы аннотаций.				
Классификация и кластеризация текстов	2	4	4	10
Интеллектуальный анализ данных: Data Mining и Text Mining. Особенности классификации и кластеризации текстов. Модели и методы автоматической классификации и кластеризации текстовой информации. Иерархические и вероятностные подходы.				
Корпусная лингвистика	2	2	2	8
Понятие корпуса. Соотношение корпуса и базы данных. Создание и применение корпусов. Обработка и преобразования корпуса текста: сегментация, лексемизация, промежуточный анализ корпуса. Разметка корпуса. Виды разметки и области их применения. Обзор существующих общедоступных корпусов.				
Машинный перевод	2	2	4	8
Стратегии машинного перевода, основанного на лингвистических правилах. Статистический машинный перевод: особенности и виды. Принципы создания статистического переводчика. Современные подходы к машинному переводу. Многоязычный машинный перевод. Стратегии перевода специализированных текстов (деловых, технических)				
Семантический анализ текста	2	4	4	10
Способы представления смысла текста. Понятие денотата, сигнификата, референта. Треугольник Фреге. Семантический анализ текста на основе семантико-синтаксических моделей управления. Разметка частей речи. Выделение именованных сущностей. Извлечение информации и отношений из текста. Извлечение информации и знаний из текстов. Подход А.И. Новикова. Денотатный граф. Моделирование предметной области средствами денотатного графа. Модель «СМЫСЛ <-> ТЕКСТ» И.А. Мельчука. Ограничения существующих семантических моделей.				
Методы машинного обучения в обработке естественного языка	2	4	4	12
Формальные методы определения автора текста. Статистические методы атрибуции. Авторский				

Наименование разделов дисциплины с кратким содержанием	Объем аудиторных занятий по видам в часах			Объем внеаудиторных занятий по видам в часах
	Л	ЛР	ПЗ	СРС
инвариант и лингвистические спектры. Применение методов кластеризации и классификации для установления авторства текстов. Методы обнаружения спама. Автоматический анализ тональности текстов и извлечение мнений из текстов. Искусственные нейронные сети как основа для лингвистических моделей. Большие лингвистические модели (LLM). Современные архитектуры ИНС для лингвистических задач: трансформеры, диффузионные модели. Применение лингвистических моделей в задачах text2image, text2speech, text2video. GPT, BERT, CLIP, Whisper.				
ИТОГО по 4-му семестру	18	24	26	72
ИТОГО по дисциплине	18	24	26	72

Тематика примерных практических занятий

№ п.п.	Наименование темы практического (семинарского) занятия
1	Обзор возможностей пакета NLTK
2	Токенизация текста средствами библиотек Python
3	Метрика TF-IDF и инструменты ее вычисления
4	Кластеризация коллекции документов

Тематика примерных лабораторных работ

№ п.п.	Наименование темы лабораторной работы
1	Сравнение морфологических процессоров для русского языка
2	Построение векторной модели коллекции документов
3	Квазиреферирование специализированного текста
4	Обзор нейросетевых лингвистических моделей

5. Организационно-педагогические условия

5.1. Образовательные технологии, используемые для формирования компетенций

Проведение лекционных занятий по дисциплине основывается на активном методе обучения, при котором учащиеся не пассивные слушатели, а активные участники занятия, отвечающие на вопросы преподавателя. Вопросы преподавателя нацелены на активизацию процессов усвоения материала, а также на развитие логического мышления. Преподаватель заранее намечает список вопросов, стимулирующих ассоциативное мышление и установление связей с ранее освоенным материалом.

Практические занятия проводятся на основе реализации метода обучения действием: определяются проблемные области, формируются группы. При проведении практических занятий преследуются следующие цели: применение знаний отдельных дисциплин и креативных методов для решения проблем и приятия решений; отработка у обучающихся навыков командной работы, межличностных коммуникаций и развитие лидерских качеств; закрепление основ теоретических знаний.

Проведение лабораторных занятий основывается на интерактивном методе обучения, при котором обучающиеся взаимодействуют не только с преподавателем, но и друг с другом. При этом доминирует активность учащихся в процессе обучения. Место преподавателя в интерактивных занятиях сводится к направлению деятельности обучающихся на достижение целей занятия.

При проведении учебных занятий используются интерактивные лекции, групповые дискуссии, ролевые игры, тренинги и анализ ситуаций и имитационных моделей.

5.2. Методические указания для обучающихся по изучению дисциплины

При изучении дисциплины обучающимся целесообразно выполнять следующие рекомендации:

1. Изучение учебной дисциплины должно вестись систематически.
2. После изучения какого-либо раздела по учебнику или конспектным материалам рекомендуется по памяти воспроизвести основные термины, определения, понятия раздела.
3. Особое внимание следует уделить выполнению отчетов по практическим занятиям, лабораторным работам и индивидуальным комплексным заданиям на самостоятельную работу.
4. Вся тематика вопросов, изучаемых самостоятельно, задается на лекциях преподавателем. Им же даются источники (в первую очередь вновь изданные в периодической научной литературе) для более детального понимания вопросов, озвученных на лекции.

6. Перечень учебно-методического и информационного обеспечения для самостоятельной работы обучающихся по дисциплине

6.1. Печатная учебно-методическая литература

№ п/п	Библиографическое описание (автор, заглавие, вид издания, место, издательство, год издания, количество страниц)	Количество экземпляров в библиотеке
1. Основная литература		
1	Доусон М. Программируем на Python : пер. с англ. Санкт-Петербург [и др.] : Питер, 2021. 414 с. 33,540 усл. печ. л.	6
2	Леонтьева Н.Н Автоматическое понимание текстов: системы, модели, ресурсы : учебное пособие для вузов. М. : Академия, 2006. 303 с.	3
2. Дополнительная литература		

2.1. Учебные и научные издания		
1	Васильев А. Н. Python на примерах. Практический курс по программированию. 3-е изд. Санкт-Петербург : Наука и техника, 2019. 428 с. 27 усл. печ. л.	6
2	Хопкрофт Д., Мотвани Р., Ульман Д. Введение в теорию автоматов, языков и вычислений : пер. с англ. 2-е изд., испр. М. : Вильямс, 2008. 527 с.	3
2.2. Периодические издания		
	Не используется	
2.3. Нормативно-технические издания		
	Не используется	
3. Методические указания для студентов по освоению дисциплины		
	Не используется	
4. Учебно-методическое обеспечение самостоятельной работы студента		
	Не используется	

6.2. Электронная учебно-методическая литература

Вид литературы	Наименование разработки	Ссылка на информационный ресурс	Доступность (сеть Интернет / локальная сеть; авторизованный / свободный доступ)
Дополнительная литература	Мельчук, И. А. Опыт теории лингвистических моделей «СМЫСЛ - ТЕКСТ» [Электронный ресурс]	https://e.lanbook.com/book/137010	локальная сеть; свободный доступ
Дополнительная литература	Митчелл, Р. Скрапинг веб-сайтов с помощью Python [Электронный ресурс]	https://e.lanbook.com/book/100903	локальная сеть; свободный доступ
Основная литература	Бонцанини, М. Анализ социальных медиа на Python. Извлекайте и анализируйте данные из всех уголков социальной паутины на Python [Электронный ресурс]	https://e.lanbook.com/book/108129	локальная сеть; свободный доступ
Основная литература	Маккинли Уэс. Python и анализ данных [Электронный ресурс]	https://www.iprbookshop.ru/88752.html	локальная сеть; свободный доступ

6.3. Лицензионное и свободно распространяемое программное обеспечение, используемое при осуществлении образовательного процесса по дисциплине

Вид ПО	Наименование ПО
Операционные системы	Debian (GNU GPL)
Офисные приложения.	LibreOffice 6.2.4. OpenSource, бесплатен.

Вид ПО	Наименование ПО
Системы управления проектами, исследованиями, разработкой, проектированием, моделированием и внедрением	Protege
Среды разработки, тестирования и отладки	PIP (The Python Package Installer) Free
Среды разработки, тестирования и отладки	PostgreSQL (PostgreSQL License)

6.4. Современные профессиональные базы данных и информационные справочные системы, используемые при осуществлении образовательного процесса по дисциплине

Наименование	Ссылка на информационный ресурс
База данных Elsevier "Freedom Collection"	https://www.elsevier.com/
База данных Scopus	https://www.scopus.com/
База данных научной электронной библиотеки (eLIBRARY.RU)	https://elibrary.ru/
Научная библиотека Пермского национального исследовательского политехнического университета	http://lib.pstu.ru/
Электронно-библиотечеая система Лань	https://e.lanbook.com/
Электронно-библиотечная система IPRbooks	http://www.iprbookshop.ru/
Информационные ресурсы Сети КонсультантПлюс	http://www.consultant.ru/
Информационно-справочная система нормативно-технической документации "Техэксперт: нормы, правила, стандарты и законодательства России"	https://техэксперт.сайт/

7. Материально-техническое обеспечение образовательного процесса по дисциплине

Вид занятий	Наименование необходимого основного оборудования и технических средств обучения	Количество единиц
Лабораторная работа	ПК с предустановленным интерпретатором Python версии 3.8 или выше и средой разработки	10
Лекция	Мультимедийный проектор	1
Практическое занятие	ПК с предустановленным интерпретатором Python версии 3.8 или выше и средой разработки	10

8. Фонд оценочных средств дисциплины

Описан в отдельном документе

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Пермский национальный исследовательский политехнический университет»

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ
для проведения промежуточной аттестации обучающихся по дисциплине «Модели и методы обработки естественного языка»
Приложение к рабочей программе дисциплины

Направление подготовки:	09.04.04 Программная инженерия
Направленность (профиль) образовательной программы:	Разработка программно-информационных систем
Квалификация выпускника:	«Магистр»
Выпускающая кафедра:	Информационные технологии и автоматизированные системы
Форма обучения:	Очная

Курс: 2 **Семестр**: 4

Трудоёмкость:
Кредитов по рабочему учебному плану: 4 ЗЕ
Часов по рабочему учебному плану: 144 ч.

Форма промежуточной аттестации:
Дифференцированный зачет: 4 семестр

Пермь 2023 г.

Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине является частью (приложением) к рабочей программе дисциплины. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине разработан в соответствии с общей частью фонда оценочных средств для проведения промежуточной аттестации основной образовательной программы, которая устанавливает систему оценивания результатов промежуточной аттестации и критерии выставления оценок. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине устанавливает формы и процедуры текущего контроля успеваемости и промежуточной аттестации обучающихся по дисциплине.

1. Перечень контролируемых результатов обучения по дисциплине, объекты оценивания и виды контроля

Согласно РПД, освоение учебного материала дисциплины запланировано в течение одного семестра (4-го семестра учебного плана) и разбито на 8 учебных модулей. В каждом модуле предусмотрены аудиторные лекционные, практические и лабораторные занятия, а также самостоятельная работа студентов. В рамках освоения учебного материала дисциплины формируется компоненты компетенций *знать, уметь, владеть*, указанные в РПД, которые выступают в качестве контролируемых результатов обучения по дисциплине (таблица 1.1).

Контроль уровня усвоенных знаний, освоенных умений и приобретенных владений осуществляется в рамках текущего, рубежного и промежуточного контроля при изучении теоретического материала, сдаче отчетов по лабораторным работам и дифференцированного зачета. Виды контроля сведены в таблицу 1.1.

Таблица 1.1. Перечень контролируемых результатов обучения по дисциплине

Контролируемые результаты обучения по дисциплине (ЗУВы)	**Вид контроля**					
	Текущий		**Промежуточный /рубежный**		**Итоговый**	
	С	**ТО**	**ОЛР**	**Т/КР**	**Диф. зачет**	
Усвоенные знания						
З.1 Знает методологии разработки программного обеспечения обработки естественного языка		ТО			ТЗ	
Освоенные умения						
У.1 Умеет применять нормативные документы, определяющие требования к оформлению программного кода информационных систем обработки естественного языка			ОЛР2 ОЛР3		КЗ	
Приобретенные владения						
В.1 Владеет навыками оценки качества и эффективности программного кода в информационных системах обработки естественного языка			ОЛР1 ОЛР4		КЗ	

С – собеседование по теме; ТО – коллоквиум (теоретический опрос); КЗ – кейс-задача (индивидуальное задание); ОЛР – отчет по лабораторной работе; Т/КР – рубежное тестирование (контрольная работа); ТВ – теоретический вопрос; ПЗ – практическое задание; КЗ – комплексное задание экзамена.

Итоговой оценкой достижения результатов обучения по дисциплине является промежуточная аттестация в виде дифференцированного зачета,

проводимая с учетом результатов текущего и рубежного контроля.

2. Виды контроля, типовые контрольные задания и шкалы оценивания результатов обучения

Текущий контроль успеваемости имеет целью обеспечение максимальной эффективности учебного процесса, управление процессом формирования заданных компетенций обучаемых, повышение мотивации к учебе и предусматривает оценивание хода освоения дисциплины. В соответствии с Положением о проведении текущего контроля успеваемости и промежуточной аттестации обучающихся по образовательным программам высшего образования – программам бакалавриата, специалитета и магистратуры в ПНИПУ предусмотрены следующие виды и периодичность текущего контроля успеваемости обучающихся:

- входной контроль, проверка исходного уровня подготовленности обучаемого и его соответствия предъявляемым требованиям для изучения данной дисциплины;
- текущий контроль усвоения материала (уровня освоения компонента «знать» заданных компетенций) на каждом групповом занятии и контроль посещаемости лекционных занятий;
- промежуточный и рубежный контроль освоения обучаемыми отдельных компонентов «знать», «уметь» заданных компетенций путем компьютерного или бланочного тестирования, контрольных опросов, контрольных работ (индивидуальных домашних заданий), защиты отчетов по лабораторным работам, рефератов, эссе и т.д.

Рубежный контроль по дисциплине проводится на следующей неделе после прохождения модуля дисциплины, а промежуточный – во время каждого контрольного мероприятия внутри модулей дисциплины;

- межсессионная аттестация, единовременное подведение итогов текущей успеваемости не менее одного раза в семестр по всем дисциплинам для каждого направления подготовки (специальности), курса, группы;
- контроль остаточных знаний.

2.1. Текущий контроль усвоения материала

Текущий контроль усвоения материала в форме собеседования или выборочного теоретического опроса студентов проводится по каждой теме. Результаты по 4-балльной шкале оценивания заносятся в книжку преподавателя и учитываются в виде интегральной оценки при проведении промежуточной аттестации.

2.2. Рубежный (промежуточный) контроль

Рубежный (промежуточный) контроль для комплексного оценивания усвоенных знаний, освоенных умений и приобретенных владений (таблица 1.1) проводится в форме защиты лабораторных работ и выполнения кейс-задач.

2.2.1. Защита лабораторных работ

Всего запланировано 4 лабораторных работы. Типовые темы лабораторных работ приведены в РПД.

Защита лабораторной работы проводится индивидуально каждым студентом

или группой студентов. Типовые шкала и критерии оценки приведены в общей части ФОС образовательной программы.

2.3. Промежуточная аттестация (итоговый контроль)

Допуск к промежуточной аттестации осуществляется по результатам текущего и рубежного контроля. Условиями допуска являются успешная сдача всех лабораторных работ и положительная интегральная оценка по результатам текущего и рубежного контроля.

2.3.1. Процедура промежуточной аттестации без дополнительного аттестационного испытания

Промежуточная аттестация проводится в форме дифференцированного зачета. Зачет по дисциплине основывается на результатах выполнения лабораторных работ и участия в семинарах по данной дисциплине.

Критерии выведения итоговой оценки за компоненты компетенций при проведении промежуточной аттестации в виде зачета приведены в общей части ФОС образовательной программы.

2.3.2. Процедура промежуточной аттестации с проведением аттестационного испытания

В отдельных случаях (например, в случае переаттестации дисциплины) промежуточная аттестация в виде зачета по дисциплине может проводиться с проведением аттестационного испытания по билетам. Билет содержит теоретические вопросы (ТВ) для проверки усвоенных знаний и комплексные задания (КЗ) для контроля уровня освоенных умений и приобретенных владений всех заявленных компетенций.

Билет формируется таким образом, чтобы в него попали вопросы и практические задания, контролирующие уровень сформированности всех заявленных компетенций.

2.4.2.1. Типовые вопросы и задания для зачета по дисциплине

Типовые вопросы для контроля усвоенных знаний:

1. Уровни языка.
2. Этапы анализа текста.
3. Основные задачи обработки естественного языка.
4. Языковые модели.
5. Методы сбора текстовых данных.
6. Стемминг. Лемматизация,
7. Векторная модель документа.
8. Оценка релевантности документа.
9. Квазиреферирование и автоматическое аннотирование документов.
10. Интеллектуальный анализ данных: Data Mining и Text Mining.
11. Модели и методы автоматической классификации и кластеризации текстовой информации.
12. Понятие корпуса текстов. Соотношение корпуса и базы данных.
13. Разметка корпуса текстов. Виды разметки и области их применения.

14. Общие стратегии машинного перевода.
15. Стратегии перевода специализированных текстов (деловых, технических).
16. Современные подходы к машинному переводу.
17. Многоязычный машинный перевод.
18. Способы представления смысла текста.
19. Статистические методы атрибуции текстов.
20. Большие лингвистические модели (LLM).

Типовые комплексные задания для контроля освоенных умений и приобретенных владений:

1. Описать процедуру построения векторной модели коллекции документов
2. Описать алгоритм кластеризации коллекции документов.
3. Предложить методику квазиреферирования специализированного текста.

2.3.2.2. Шкалы оценивания результатов обучения на зачете

Оценка результатов обучения по дисциплине в форме уровня сформированности компонентов *знать, уметь, владеть* заявленных компетенций проводится по 4-х балльной шкале оценивания.

Типовые шкала и критерии оценки результатов обучения при сдаче зачета для компонентов *знать, уметь и владеть* приведены в общей части ФОС образовательной программы.

3. Критерии оценивания уровня сформированности компонентов и компетенций

3.1. Оценка уровня сформированности компонентов компетенций

При оценке уровня сформированности компетенций в рамках выборочного контроля при экзамене считается, что *полученная оценка за компонент проверяемой в билете компетенции обобщается на соответствующий компонент всех компетенций, формируемых в рамках данной учебной дисциплины*.

Типовые критерии и шкалы оценивания уровня сформированности компонентов компетенций приведены в общей части ФОС образовательной программы.

3.2. Оценка уровня сформированности компетенций

Общая оценка уровня сформированности всех компетенций проводится путем агрегирования оценок, полученных студентом за каждый компонент формируемых компетенций, с учетом результатов текущего и рубежного контроля в виде интегральной оценки по 4-х балльной шкале. Все результаты контроля заносятся в оценочный лист и заполняются преподавателем по итогам промежуточной аттестации.

Форма оценочного листа и требования к его заполнению приведены в общей части ФОС образовательной программы.

При формировании итоговой оценки промежуточной аттестации в виде экзамена используются типовые критерии, приведенные в общей части ФОС образовательной программы.