

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Нижегородский государственный
университет им. Н.И. Лобачевского»

Институт информационных технологий, математики и механики
(факультет / институт / филиал)

УТВЕРЖДЕНО
решением Ученого совета ННГУ
протокол от
«30» ноября 2022 г. № 13

Рабочая программа дисциплины

Обработка естественных языков

(наименование дисциплины (модуля))

Уровень высшего образования

Магистратура

(бакалавриат / магистратура / специалитет)

Направление подготовки / специальность

09.04.04 Программная инженерия

(указывается код и наименование направления подготовки / специальности)

Направленность образовательной программы

Технологии цифровой трансформации

(указывается профиль / магистерская программа / специализация)

Форма обучения

Очная

(очная / очно-заочная / заочная)

Нижний Новгород

2023

1. Место дисциплины (модуля) в структуре ОПОП

Дисциплина «Б1.В.ДВ.02.02,Обработка естественных языков» относится к дисциплинам по выбору части, формируемой участниками образовательных отношений Блока 1 «Дисциплины (модули)» направления подготовки 09.04.04 «Программная инженерия» профиля подготовки «Технологии цифровой трансформации». Дисциплина преподается в 3 семестре. Трудоемкость дисциплины составляет 3 зачетные единицы, 108 час., зачет.

№ варианта	Место дисциплины в учебном плане образовательной программы	Стандартный текст для автоматического заполнения в конструкторе РПД
2	Блок 1. Дисциплины (модули) Часть, формируемая участниками образовательных отношений	Дисциплина «Б1.В.ДВ.02.02,Обработка естественных языков» относится к части ООП направления подготовки 09.04.04 «Программная инженерия», формируемой участниками образовательных отношений

2. Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями и индикаторами достижения компетенций)

№ п/п	Код компетенции	Содержание компетенции	Планируемые результаты обучения	Наименование оценочного средства
1	ПК-11	Владеет методами организационного и технологического обеспечения проектирования и дизайна ИС	ПК-11.1. Знает инструменты и методы проектирования и дизайна ИС	<i>Собеседование</i>
			ПК-11.2. Умеет проводить обеспечение соответствия проектирования и дизайна ИС принятым в организации или проекте стандартам и технологиям	<i>Собеседование Практическая работа</i>
			ПК-11.3. Имеет практический опыт верификации структуры программного кода ИС	<i>Собеседование Практическая работа</i>

3. Структура и содержание дисциплины

3.1. Трудоемкость дисциплины

Объем дисциплины (модуля) составляет

3 зачетные единицы, всего 108 час., из которых

65 час. составляет **контактная** работа обучающегося с преподавателем:

32 часа занятия лекционного типа,

32 часа. занятия семинарского типа (семинары, лабораторные работы и т.п.),

1 час. мероприятия промежуточной аттестации

43 час. составляет **самостоятельная** работа обучающегося.

3.2. Содержание дисциплины

Наименование и краткое содержание разделов и тем дисциплины	Всего (часы) Оч н а я	В том числе				
		Контактная работа (работа во взаимодействии с преподавателем), часы. Из них				Самостоятельная работа обучающегося, часы
		Занятия лекционного типа Очная	Занятия семинарского типа Очная	Занятия лабораторного типа	Всего	
Введение в предмет. Основные задачи и методы. Автоматическая обработка текстов (АОТ). Сфера использования. Проблема неоднозначности в автоматической обработке текстов (лексическая, синтаксическая, семантическая неоднозначности, неоднозначности на уровне дискурса, на уровне прагматики и др.). Морфологическая разметка. Синтаксический разбор. Семантический анализ.	8	2		2	4	4
Компьютерная морфология. Морфологический анализ. Словарный и предиктивный морфологический анализ. Лексическая неоднозначность. Инструменты для морфологического анализа и методика их работы (АОТ, PyMorphy, MyStem, NLTK). Языковая модель. Цепь Маркова, <i>n</i> -граммы. Задача определения части речи. Статистические методы определения части речи. Частеречевая разметка на базе скрытых Марковских цепей и	14	2		2	4	10

алгоритм Витерби.						
Исправление опечаток. Расстояние Левенштейна, расстояние Левенштейна–Дамерау. Подсчет расстояний Левенштейна. Инструментарий для исправления опечаток. Морфологическая классификация естественных языков. Лингвистическая типология.	14	2		2	4	10
Синтаксический анализ в естественном языке. Синтаксическая неоднозначность. Подходы к описанию синтаксиса в естественном языке. Иерархия Хомского. Задача синтаксического разбора. Грамматика зависимостей. Методы и алгоритмы синтаксического разбора в контексте грамматики зависимостей. Возможности и ограничения грамматики зависимостей.	14	2		2	4	10
Контекстно-свободные грамматики (КС-грамматики). Методы и алгоритмы синтаксического разбора в контексте КС-грамматик. Возможности и ограничения КС-грамматики. КС-грамматика как дополнение грамматики зависимостей.	14	2		2	4	10
Статистические методы синтаксического анализа. Оценка точности синтаксического анализа. Понятие проективности. SyntaxNet.	14	2		2	4	10
Семантический анализ. Формальные методы семантического анализа. Понятие онтологии. Модели представления знаний в компьютерной семантике. Онтологические ресурсы и компьютерные тезаурусы. Ресурсы WordNet, FrameNet. Тезаурусы для русского языка.	14	2		2	4	10
Дистрибутивная семантика. Word2Vec. Алгоритмы CBOW и Модель Skip-gram, GloVe. Исследование свойств предобученной модели Skip-gram модели, обучение своей.	14	2		2	4	10
Текущий контроль (КСР)	2				2	
Промежуточная аттестация – зачет	36				36	
Итого	144	16		16	70	74

Практические занятия (семинарские занятия /лабораторные работы) организуются, в том числе в форме практической подготовки, которая предусматривает участие обучающихся в выполнении отдельных элементов работ, связанных с будущей профессиональной деятельностью.

Практическая подготовка предусматривает: изучение методических материалов, подготовку к вопросам к экзамену, выполнение практических заданий.

На проведение практических занятий (семинарских занятий /лабораторных работ) в форме практической подготовки отводится 32 часа.

Практическая подготовка направлена на формирование и развитие:

- практических навыков в соответствии с профилем ОП: создание и сопровождение архитектуры программных средств, разработка и тестирование программного обеспечения;
- компетенций – ПК-11.

Текущий контроль успеваемости реализуется в формах опросов на занятиях семинарского типа

Промежуточная аттестация проходит в традиционных формах (зачет)

4. Учебно-методическое обеспечение самостоятельной работы обучающихся

Самостоятельная работа заключается в чтении литературы из списка основной литературы и решения практических заданий. По ходу выполнения самостоятельной работы возможны консультации с преподавателем посредством электронной почты и социальных сетей.

Контрольные вопросы и задания для проведения текущего контроля и промежуточной аттестации по итогам освоения дисциплины приведены в п. 5.2.

5. Фонд оценочных средств для промежуточной аттестации по дисциплине (модулю),

включающий:

5.1. Описание шкал оценивания результатов обучения по дисциплине

Уровень сформированности компетенций (индикатора достижения компетенций)	Шкала оценивания сформированности компетенций						
	плохо	неудовлетворительно	удовлетворительно	хорошо	очень хорошо	отлично	превосходно
	Не зачтено		Зачтено				
<u>Знания</u>	Отсутствие знаний теоретического материала. Невозможность оценить полноту знаний вследствие	Уровень знаний ниже минимальных требований. Имели место грубые ошибки.	Минимально допустимый уровень знаний. Допущено много негрубых ошибок.	Уровень знаний в объеме, соответствующем программе подготовки. Допущено несколько негрубых ошибок	Уровень знаний в объеме, соответствующем программе подготовки. Допущено несколько негрубых ошибок	Уровень знаний в объеме, соответствующем программе подготовки, без ошибок.	Уровень знаний в объеме, превышающем программные требования

	отказа обучающегося от ответа						
<u>Умения</u>	Отсутствие минимальных умений. Невозможность оценить наличие умений вследствие отказа обучающегося от ответа	При решении стандартных задач не продемонстрированы основные умения. Имели место грубые ошибки.	Продemonстрированы основные умения. Решены типовые задачи с негрубыми ошибками. Выполнены все задания, но не в полном объеме.	Продemonстрированы все основные умения. Решены все основные задачи с негрубыми ошибками. Выполнены все задания, в полном объеме, но некоторые с недочетами.	Продemonстрированы все основные умения. Решены все основные задачи. Выполнены все задания, в полном объеме, но некоторые с недочетами.	Продemonстрированы все основные умения, решены все основные задачи с отдельными несущественными недочетами, выполнены все задания в полном объеме.	Продemonстрированы все основные умения, решены все основные задачи. Выполнены все задания в полном объеме без недочетов
<u>Навыки</u>	Отсутствие владения материалом. Невозможность оценить наличие навыков вследствие отказа обучающегося от ответа	При решении стандартных задач не продемонстрированы базовые навыки. Имели место грубые ошибки.	Имеется минимальный набор навыков для решения стандартных задач с некоторыми недочетами.	Продemonстрированы базовые навыки при решении стандартных задач с некоторыми недочетами	Продemonстрированы базовые навыки при решении стандартных задач без ошибок и недочетов.	Продemonстрированы навыки при решении нестандартных задач без ошибок и недочетов.	Продemonстрирован творческий подход к решению нестандартных задач.

Шкала оценки при промежуточной аттестации

Оценка		Уровень подготовки
зачтено	Превосходно	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «превосходно»
	Отлично	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «отлично», при этом хотя бы одна компетенция сформирована на уровне «отлично»
	Очень хорошо	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «очень хорошо», при этом хотя бы одна компетенция сформирована на уровне «очень хорошо»
	Хорошо	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «хорошо», при этом хотя бы одна компетенция сформирована на уровне «хорошо»
	Удовлетворительно	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «удовлетворительно», при этом хотя бы одна компетенция сформирована на уровне «удовлетворительно»
не зачтено	Неудовлетворительно	Хотя бы одна компетенция сформирована на уровне «неудовлетворительно», ни одна из компетенций не сформирована на уровне «плохо»
	Плохо	Хотя бы одна компетенция сформирована на уровне «плохо»

5.2. Типовые контрольные задания или иные материалы, необходимые для оценки результатов обучения

5.2.1 Контрольные вопросы (ПК-11.1)

<i>Вопросы</i>	<i>Код формируемой компетенции</i>
1. Сложность АОТ. Неоднозначность при обработке естественного языка. Уровни неоднозначности.	ПК-11.1
2. Основные задачи АОТ	ПК-11.1
3. Предмет компьютерной морфологии. Морфологический анализ. Словарный и предиктивный морфологический анализ.	ПК-11.1
4. Подходы к определению грамматического значения несловарных слов. Лексическая неоднозначность в морфологическом анализе.	ПК-11.1
5. Морфологический анализ на базе правил. Инструменты для морфологического анализа (АОТ, PyMorphy, MyStem) и методика их работы.	ПК-11.1
6. Задача частеречевой разметки. Статистическая частеречевая разметка.	ПК-11.1
7. Понятие скрытой Марковской модели (НММ). Алгоритм Витерби. Использование алгоритма Витерби для решения задачи частеречевой разметки. Учет незнакомых слов при статистическом подходе к чатеречевой разметке.	ПК-11.1
8. Исправление опечаток. Расстояние Левенштейна, расстояние Левенштейна-Дамерау. Подсчет расстояний Левенштейна. Инструментарий.	ПК-11.1
9. Морфологическая классификация языков. Примеры.	ПК-11.1
10. Синтаксический анализ в естественном языке. Проблематика. Синтаксическая неоднозначность. Подходы к описанию синтаксиса естественного языка. Иерархия Хомского.	ПК-11.1
11. Грамматика зависимостей. Методы. Проблемы (придаточные предложения, и т.д.). Недостаточность ГЗ. Понятие грамматики непосредственно составляющих. Алгоритмы парсинга грамматики НС.	ПК-11.1
12. Грамматика непосредственно составляющих. Алгоритмы. Проблема неоднозначности и комбинаторного взрыва.	ПК-11.1
13. Алгоритмы статистического парсинга. КС-грамматики. Вероятностные КС-грамматики. Алгоритм СКУ. Оценка качества синтаксического разбора.	ПК-11.1
14. Лексикализация. Dependency Parsing. Проективность и непроективность при парсинге. Оценка качества синтаксического разбора ГЗ. SyntaxNet.	ПК-11.1
15. Семантический анализ. Модели представления знаний в компьютерной семантике (сетевые модели, концептуальные графы, фреймы и сценарии, современные подходы).	ПК-11.1
16. Понятие формальной онтологии. Онтологические ресурсы.	ПК-11.1
17. Компьютерные тезаурусы. WordNet, FrameNet. Тезаурусы для русского языка.	ПК-11.1

18. Дистрибутивная семантика. Понятие дистрибутивной семантики. Классические count-based подходы к дистрибутивной семантике. Векторное представление слова.	ПК-11.1
19. Предиктивные подходы в дистрибутивной семантике. Word2vec. Алгоритмы CBOW и skip-gram. Deep learning и word2vec.	ПК-11.1
20. Word2vec. Подход Миколова к ускорению Word2Vec (Hierarchical SoftMax и Negative Sampling). Лингвистические особенности и инструментарий.	ПК-11.1

5.2.2. Типовые задания/задачи для оценки сформированности компетенции ПК-11.2, ПК-11.3

Задания	Компетенция																
<p><u>Задание 1</u></p> <ul style="list-style-type: none"> - Выбрать язык из корпуса проекта Universal Dependencies (http://universaldependencies.org/). - Выполнить преобразование формата CoNLL-u - Разработать PoS-теггер на базе скрытой Марковской цепи и алгоритма Витерби. - Оценить точность частеречевой разметки. <p><u>Задание 2.</u></p> <ul style="list-style-type: none"> - Выбрать 2 языка из корпуса проекта Universal Dependencies (http://universaldependencies.org/). Языки должны относиться к разным семействам языков (с точки зрения лингвистической типологии). - Обучить синтаксический анализатор SyntaxNet, получив для выбранных языков модели для проведения морфосинтаксического анализа. - Провести тестирование полученных моделей на тренировочных корпусах выбранных языков. - Собрать статистику по тестовым корпусам, проанализировав ошибки частеречевой разметки, порождаемые морфосинтаксическим анализатором. <p><u>Задание 3.</u></p> <p>Используя WordNet применить онтологическую модель для анализа семантики текста.</p> <p><u>Задание 4.</u></p> <p>Применить алгоритм SKY для получения дерева синтаксического разбора по заданному предложению и грамматике.</p> <p>Пример</p> <p>Задана формальная грамматика :</p> <table> <tr> <td>$S \rightarrow NP VP$</td><td>$Det \rightarrow that this the a$</td></tr> <tr> <td>$S \rightarrow Aux NP VP$</td><td>$Noun \rightarrow book flight meal money tickets$</td></tr> <tr> <td>$S \rightarrow VP$</td><td>$Verb \rightarrow book include prefer$</td></tr> <tr> <td>$NP \rightarrow Pronoun$</td><td>$Pronoun \rightarrow I she me$</td></tr> <tr> <td>$NP \rightarrow Proper-Noun$</td><td>$Proper-Noun \rightarrow Houston NWA$</td></tr> <tr> <td>$NP \rightarrow Det Nominal$</td><td>$Aux \rightarrow does$</td></tr> <tr> <td>$Nominal \rightarrow Noun$</td><td>$Preposition \rightarrow from to on near through$</td></tr> <tr> <td>$Nominal \rightarrow Nominal Noun$</td><td></td></tr> </table>	$S \rightarrow NP VP$	$Det \rightarrow that this the a$	$S \rightarrow Aux NP VP$	$Noun \rightarrow book flight meal money tickets$	$S \rightarrow VP$	$Verb \rightarrow book include prefer$	$NP \rightarrow Pronoun$	$Pronoun \rightarrow I she me$	$NP \rightarrow Proper-Noun$	$Proper-Noun \rightarrow Houston NWA$	$NP \rightarrow Det Nominal$	$Aux \rightarrow does$	$Nominal \rightarrow Noun$	$Preposition \rightarrow from to on near through$	$Nominal \rightarrow Nominal Noun$		ПК-11.2
$S \rightarrow NP VP$	$Det \rightarrow that this the a$																
$S \rightarrow Aux NP VP$	$Noun \rightarrow book flight meal money tickets$																
$S \rightarrow VP$	$Verb \rightarrow book include prefer$																
$NP \rightarrow Pronoun$	$Pronoun \rightarrow I she me$																
$NP \rightarrow Proper-Noun$	$Proper-Noun \rightarrow Houston NWA$																
$NP \rightarrow Det Nominal$	$Aux \rightarrow does$																
$Nominal \rightarrow Noun$	$Preposition \rightarrow from to on near through$																
$Nominal \rightarrow Nominal Noun$																	

<p>Nominal → Nominal PP VP → Verb VP → Verb NP VP → Verb NP PP VP → Verb PP VP → VP PP PP → Preposition NP</p> <p>Привести к нормальной форме Хомского и применить алгоритм СКУ для построения дерева составляющих для строки I book the tickets to the Houston.</p> <p><u>Задача 5.</u> Подсчитать расстояния Левенштейна и Левенштейна-Дамерау для заданных строк, например, Кошка съела мышку. Мышка съела кошку.</p>	
<p><u>Задание 6</u> - Выбрать 2 языка из корпуса проекта Universal Dependencies (http://universaldependencies.org/). Языки должны относиться к разным семейства языков (с точки зрения лингвистической типологии). - Обучить синтаксический анализатор SyntaxNet, получив для выбранных языков модели для проведения морфосинтаксического анализа. - Провести тестирование полученной моделей на тренировочных корпусах выбранных языков. - Собрать статистику по тестовым корпусам, проанализировав ошибки частеречевой разметки, порождаемые морфосинтаксическим анализатором.</p> <p><u>Задание 7.</u> Применить модель word2vec для анализа близости семантики двух слов. Обучить свою модель Skip-gram.</p> <p><u>Задание 8.</u> Используя WordNet применить онтологическую модель для анализа семантики текста.</p> <p><u>Задание 9.</u> Подсчитать расстояния Левенштейна и Левенштейна-Дамерау для заданных строк, например, Дракон съел собаку. Собака подавилась драконом.</p>	ПК-11.3

6. Учебно-методическое и информационное обеспечение дисциплины

а) Основная литература

1. Добров Б., Иванов В., Лукашевич Н., Соловьев В. Онтологии и тезаурусы: модели, инструменты, приложения // Интернет университет информационных технологий.
<http://www.intuit.ru/studies/courses/1078/270/info>

б) Дополнительная литература

Афонин В., Макушкин В. Интеллектуальные робототехнические системы: Информация // Интернет университет информационных технологий.
<http://www.intuit.ru/studies/courses/46/46/info>

в) программное обеспечение и Интернет-ресурсы

Для успешного освоения дисциплины, студент использует следующие программные средства:

- Python 3.4 или R
- Библиотеки: scikit-learn, NLTK, gensim, tensorflow.
- NLPub каталог лингвистических ресурсов

7. Материально-техническое обеспечение дисциплины

Помещения представляют собой учебные аудитории для проведения учебных занятий, предусмотренных программой, оснащенные оборудованием и техническими средствами обучения: компьютерный класс, проектор, экран.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечены доступом в электронную информационно-образовательную среду.

Учебная и научная литература, учебно-методические материалы, представленные в библиотечном фонде, в электронных библиотеках и на кафедре математического обеспечения и суперкомпьютерных технологий.

Программа составлена в соответствии с требованиями ОС ВО ННГУ с учетом рекомендаций ФГОС ВО по направлению подготовки 090404 Программная инженерия

Автор: д.ф.-м.н, профессор кафедры АГДМ Золотых Н.Ю.

Рецензент: д.т.н., профессор кафедры ИАНИ Старостин Н.В.

Заведующий кафедрой: д.ф.-м.н, заведующий кафедрой АГДМ, Золотых Н.Ю.

Программа одобрена на заседании методической комиссии института информационных технологий, математики и механики от 30 ноября 2022 года, протокол № 3.