

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

УТВЕРЖДЕНО

решением президиума Ученого совета ННГУ

протокол № 1 от 16.01.2024 г.

Рабочая программа дисциплины

Обработка естественных языков

Уровень высшего образования

Магистратура

Направление подготовки / специальность

01.04.02 - Прикладная математика и информатика

Направленность образовательной программы

Анализ данных в прикладных областях

Форма обучения

очная

г. Нижний Новгород

2024 год начала подготовки

1. Место дисциплины в структуре ОПОП

Дисциплина Б1.В.ДВ.04.02 Обработка естественных языков относится к части, формируемой участниками образовательных отношений образовательной программы.

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

Формируемые компетенции (код, содержание компетенции)	Планируемые результаты обучения по дисциплине (модулю), в соответствии с индикатором достижения компетенции		Наименование оценочного средства	
	Индикатор достижения компетенции (код, содержание индикатора)	Результаты обучения по дисциплине	Для текущего контроля успеваемости	Для промежуточной аттестации
ПК-11: Способен разрабатывать и анализировать концептуальные и теоретические модели решаемых задач производственно-технологической деятельности	<p>ПК-11.1: Знает методы разработки и анализа концептуальных и теоретических моделей решаемых производственно-технологических задач</p> <p>ПК-11.2: Умеет применять методы разработки и анализа концептуальных и теоретических моделей решаемых производственно-технологических задач</p> <p>ПК-11.3: Имеет навыки применения методов разработки и анализа концептуальных и теоретических моделей решаемых производственно-технологических задач</p>	<p>ПК-11.1: Знать методы разработки и анализа концептуальных и теоретических моделей решаемых задач обработки естественных языков.</p> <p>ПК-11.2: Уметь решать основные задачи теории систем линейных неравенств и машинного обучения и применять методы разработки и анализа концептуальных и теоретических моделей.</p> <p>ПК-11.3: Владеть способностями осуществлять научное руководство коллективом специалистов, создающих алгоритмы и их программные реализации решения задач автоматической обработки текстов.</p>	Задания	Зачёт: Задания
ПК-4: Способен разрабатывать и анализировать концептуальные и теоретические модели решаемых научных проблем и задач	<p>ПК-4.1: Знает методы разработки и анализа концептуальных и теоретических моделей решаемых научных проблем и задач</p> <p>ПК-4.2: Умеет применять методы разработки и</p>	<p>ПК-4.1: Знать типовые методы применения обработки естественных языков при разработке и анализе концептуальных и теоретических моделей решаемых научных проблем и</p>	Задания	Зачёт: Контрольные вопросы

	<p>анализа концептуальных и теоретических моделей решаемых научных проблем и задач</p> <p>ПК-4.3: Имеет навыки применения методов разработки и анализа концептуальных и теоретических моделей решаемых научных проблем и задач</p>	<p>задач.</p> <p>ПК-4.2: Уметь работать с современными лингвистическими ресурсами (корпусами OpenCorpora, размеченными корпусами ГИКРЯ, семантическим корпусом и т.д.).</p> <p>ПК-4.3: Владеть практическим опытом разработки и применения системного и прикладного программного обеспечения для решения задач обработки естественных языков.</p>		
--	--	---	--	--

3. Структура и содержание дисциплины

3.1 Трудоемкость дисциплины

	очная
Общая трудоемкость, з.е.	4
Часов по учебному плану	144
в том числе	
аудиторные занятия (контактная работа):	
- занятия лекционного типа	32
- занятия семинарского типа (практические занятия / лабораторные работы)	32
- КСР	1
самостоятельная работа	79
Промежуточная аттестация	0
	Зачёт

3.2. Содержание дисциплины

(структурированное по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий)

Наименование разделов и тем дисциплины	Всего (часы)	в том числе			Самостоятельная работа обучающегося, часы
		Контактная работа (работа во взаимодействии с преподавателем), часы из них			
		Занятия лекционного типа	Занятия семинарского типа (практические занятия/ лабора торные	Всего	

			работы), часы		
	о ф о	о ф о	о ф о	о ф о	о ф о
Основные задачи и методы	35	8	8	16	19
Синтаксический анализ в естественном языке	36	8	8	16	20
Контекстно-свободные грамматики	36	8	8	16	20
Семантический анализ	36	8	8	16	20
Аттестация	0				
КСР	1			1	
Итого	144	32	32	65	79

Содержание разделов и тем дисциплины

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. Компьютерная морфология. Морфологический анализ. Словарный и предиктивный морфологический анализ. Лексическая неоднозначность. Инструменты для морфологического анализа и методика их работы (АОТ, PyMorphu, MyStem, NLTK). Языковая модель. Цепь Маркова, n-граммы. Задача определения части речи. Статистические методы определения части речи. Частеречевая разметка на базе скрытых Марковских цепей и алгоритм Витерби. Исправление опечаток. Расстояние Левенштейна, расстояние Левенштейна–Дамерау. Подсчет расстояний Левенштейна. Инструментарий для исправления опечаток. Морфологическая классификация естественных языков. Лингвистическая типология. 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. 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. Статистические методы синтаксического анализа. Оценка точности синтаксического анализа. Понятие проективности. SyntaxNet. 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. Дистрибутивная семантика. Word2Vec. Алгоритмы CBOW и Модель Skip-gram, GloVe. Исследование свойств предобученной модели Skip-gram модели, обучение своей.

4. Учебно-методическое обеспечение самостоятельной работы обучающихся

Самостоятельная работа обучающихся включает в себя подготовку к контрольным вопросам и заданиям для текущего контроля и промежуточной аттестации по итогам освоения дисциплины приведенным в п. 5.

При выполнении самостоятельной работы студентам рекомендуется использовать конспекты лекций, а также рекомендуемую в литературу:

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>

5. Фонд оценочных средств для текущего контроля успеваемости и промежуточной аттестации по дисциплине (модулю)

5.1 Типовые задания, необходимые для оценки результатов обучения при проведении текущего контроля успеваемости с указанием критериев их оценивания:

5.1.1 Типовые задания (оценочное средство - Задания) для оценки сформированности компетенции ПК-11:

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 полученной модели на тренировочных корпусах выбранных языков.
- Собрать статистику по тестовым корпусам, проанализировав ошибки частеречевой разметки, порождаемые морфосинтаксическим анализатором.

5.1.2 Типовые задания (оценочное средство - Задания) для оценки сформированности компетенции ПК-4:

Применить алгоритм СКУ для получения дерева синтаксического разбора по заданному предложению и грамматике.

Пример

Задана формальная грамматика :

$S \rightarrow NP VP$	$Det \rightarrow that \mid this \mid the \mid a$
$S \rightarrow Aux NP VP$	$Noun \rightarrow book \mid flight \mid meal \mid money \mid tickets$
$S \rightarrow VP$	$Verb \rightarrow book \mid include \mid prefer$
$NP \rightarrow Pronoun$	$Pronoun \rightarrow I \mid she \mid me$
$NP \rightarrow Proper-Noun$	$Proper-Noun \rightarrow Houston \mid NWA$
$NP \rightarrow Det Nominal$	$Aux \rightarrow does$
$Nominal \rightarrow Noun$	$Preposition \rightarrow from \mid to \mid on \mid near \mid through$

Nominal → Nominal Noun

Nominal → Nominal PP

VP → Verb

VP → Verb NP

VP → Verb NP PP

VP → Verb PP

VP → VP PP

PP → Preposition NP

Привести к нормальной форме Хомского и применить алгоритм СКУ для построения дерева составляющих для строки

I book the tickets to the Houston.

Критерии оценивания (оценочное средство - Задания)

Оценка	Критерии оценивания
зачтено	Работа выполнена в полном объеме и в срок, результаты работы алгоритма корректные на тестовых примерах, результаты работы представлены преподавателю.
не зачтено	Работа не выполнена или выполнена не в полном объеме (программа работает некорректно на тестовых примерах, результаты работы не представлены преподавателю).

5.2. Описание шкал оценивания результатов обучения по дисциплине при промежуточной аттестации

Шкала оценивания сформированности компетенций

Уровень сформированности компетенций (индикатора достижения компетенций)	плохо	неудовлетворительно	удовлетворительно	хорошо	очень хорошо	отлично	превосходно
	не зачтено		зачтено				
<u>Знания</u>	Отсутствие знаний теоретического материала. Невозможность оценить полноту знаний	Уровень знаний ниже минимальных требований. Имели место грубые ошибки	Минимально допустимый уровень знаний. Допущено много	Уровень знаний в объеме, соответствующем программе подготовки	Уровень знаний в объеме, соответствующем программе подготовки	Уровень знаний в объеме, соответствующем программе подготовки	Уровень знаний в объеме, превышающем программу подготовки.

	вследствие отказа обучающегося от ответа		негрубых ошибок	. Допущено несколько негрубых ошибок	. Допущено несколько несущественных ошибок	и. Ошибок нет.	
<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	Продемонстрированы основные умения. Решены типовые задачи с негрубыми ошибками. Выполнены все задания, но не в полном объеме	Продемонстрированы все основные умения. Решены все основные задачи с негрубыми ошибками. Выполнены все задания в полном объеме, но некоторые с недочетами	Продемонстрированы все основные умения. Решены все основные задачи. Выполнены все задания в полном объеме, но некоторые с недочетами	Продемонстрированы все основные умения. Решены все основные задачи с отдельным и несущественными недочетами, выполнены все задания в полном объеме	Продемонстрированы все основные умения. Решены все основные задачи. Выполнены все задания, в полном объеме без недочетов
<u>Навыки</u>	Отсутствие базовых навыков. 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 некоторым и недочетами	Продемонстрированы базовые навыки при решении стандартных задач с некоторым и недочетами	Продемонстрированы базовые навыки при решении стандартных задач без ошибок и недочетов	Продемонстрированы навыки при решении нестандартных задач без ошибок и недочетов	Продемонстрирован творческий подход к решению нестандартных задач

Шкала оценивания при промежуточной аттестации

Оценка		Уровень подготовки
зачтено	превосходно	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «превосходно», продемонстрированы знания, умения, владения по соответствующим компетенциям на уровне выше предусмотренного программой
	отлично	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «отлично».
	очень хорошо	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «очень хорошо»
	хорошо	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «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».
	плохо	Хотя бы одна компетенция сформирована на уровне «плохо»

5.3 Типовые контрольные задания или иные материалы, необходимые для оценки результатов обучения на промежуточной аттестации с указанием критериев их оценивания:

5.3.1 Типовые задания (оценочное средство - Задания) для оценки сформированности компетенции ПК-11

Задание 1

- 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.

Критерии оценивания (оценочное средство - Задания)

Оценка	Критерии оценивания
зачтено	Работа выполнена в полном объеме и в срок, результаты работы алгоритма корректные на тестовых примерах, результаты работы представлены преподавателю.
не зачтено	Работа не выполнена или выполнена не в полном объеме (программа работает некорректно на тестовых примерах, результаты работы не представлены преподавателю).

5.3.2 Типовые задания (оценочное средство - Контрольные вопросы) для оценки сформированности компетенции ПК-4

1. Сложность АОТ. Неоднозначность при обработке естественного языка. Уровни неоднозначности.
1. Основные задачи АОТ
1. Предмет компьютерной морфологии. Морфологический анализ. Словарный и предиктивный морфологический анализ.
1. Подходы к определению грамматического значения несловарных слов. Лексическая неоднозначность в морфологическом анализе.
1. Морфологический анализ на базе правил. Инструменты для морфологического анализа (АОТ, PyMorphu, MyStem) и методика их работы.
1. Задача частеречевой разметки. Статистическая частеречевая разметка.

1. Понятие скрытой Марковской модели (НММ). Алгоритм Витерби. Использование алгоритма Витерби для решения задачи частеречевой разметки. Учет незнакомых слов при статистическом подходе к чатеречевой разметке.
1. Исправление опечаток. Расстояние Левенштейна, расстояние Левенштейна-Дамерау. Подсчет расстояний Левенштейна. Инструментарий.
1. Морфологическая классификация языков. Примеры.
1. Синтаксический анализ в естественном языке. Проблематика. Синтаксическая неоднозначность. Подходы к описанию синтаксиса естественного языка. Иерархия Хомского.
1. Грамматика зависимостей. Методы. Проблемы (придаточные предложения, и т.д.). Недостаточность ГЗ. Понятие грамматики непосредственно составляющих. Алгоритмы парсинга грамматики НС.
1. Грамматика непосредственно составляющих. Алгоритмы. Проблема неоднозначности и комбинаторного взрыва.
1. Алгоритмы статистического парсинга. КС-грамматики. Вероятностные КС-грамматики. Алгоритм СКУ. Оценка качества синтаксического разбора.
1. Лексикализация. Dependency Parsing. Проективность и непроективность при парсинге. Оценка качества синтаксического разбора ГЗ. SyntaxNet.
1. Семантический анализ. Модели представления знаний в компьютерной семантике (сетевые модели, концептуальные графы, фреймы и сценарии, современные подходы).
1. Понятие формальной онтологии. Онтологические ресурсы.
1. Компьютерные тезаурусы. WordNet, FrameNet. Тезаурусы для русского языка.
1. Дистрибутивная семантика. Понятие дистрибутивной семантики. Классические count-based подходы к дистрибутивной семантике. Векторное представление слова.
1. Предиктивные подходы в дистрибутивной семантике. Word2vec. Алгоритмы CBOW и skip-gram. Deep learning и word2vec.
1. Word2vec. Подход Миколова к ускорению Word2Vec (Hierarchical SoftMax и Negative Sampling). Лингвистические особенности и инструментарий.

Критерии оценивания (оценочное средство - Контрольные вопросы)

Оценка	Критерии оценивания
зачтено	Владение основным и дополнительным материалом достаточное или с незначительными ошибками и погрешностями
не	владение материалом, необходимым по данному предмету, недостаточно. Работу за время

Оценка	Критерии оценивания
зачтено	семестра можно оценить как неудовлетворительную

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

Основная литература:

1. Мантусов А. Б. Обработка естественного языка с использованием языка программирования Python : учебное пособие. Ч. 1. Обработка естественного языка с использованием языка программирования Python. В 2 ч. Ч. 1 / Мантусов А. Б. - Элиста : КГУ, 2022. - 56 с. - Книга из коллекции КГУ - Информатика., <https://e-lib.unn.ru/MegaPro/UserEntry?Action=FindDocs&ids=885562&idb=0>.

Дополнительная литература:

1. Афраимович Л. Г. Тестовые задачи для самостоятельной подготовки по курсу «Теория автоматов и формальные грамматики» : учебно-методическое пособие / Афраимович Л. Г. - Нижний Новгород : ННГУ им. Н. И. Лобачевского, 2011. - 32 с. - Рекомендовано методической комиссией факультета ВМК для студентов ННГУ, обучающихся по направлению подготовки 230700 «Прикладная информатика». - Библиогр.: доступна в карточке книги, на сайте ЭБС Лань. - Книга из коллекции ННГУ им. Н. И. Лобачевского - Инженерно-технические науки., <https://e-lib.unn.ru/MegaPro/UserEntry?Action=FindDocs&ids=730317&idb=0>.

2. Миронов С. В. Формальные языки и грамматики : учебное пособие для студентов факультета компьютерных наук и информационных технологий / Миронов С. В. - Саратов : СГУ, 2019. - 80 с. - Библиогр.: доступна в карточке книги, на сайте ЭБС Лань. - Книга из коллекции СГУ - Информатика. - ISBN 978-5-292-04612-7., <https://e-lib.unn.ru/MegaPro/UserEntry?Action=FindDocs&ids=728644&idb=0>.

Программное обеспечение и Интернет-ресурсы (в соответствии с содержанием дисциплины):

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

7. Материально-техническое обеспечение дисциплины (модуля)

Учебные аудитории для проведения учебных занятий, предусмотренных образовательной программой, оснащены мультимедийным оборудованием (проектор, экран), техническими средствами обучения.

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

Программа составлена в соответствии с требованиями ОС ННГУ по направлению подготовки/специальности 01.04.02 - Прикладная математика и информатика.

Автор(ы): Золотых Николай Юрьевич, доктор физико-математических наук, доцент.

Заведующий кафедрой: Золотых Николай Юрьевич, доктор физико-математических наук.

Программа одобрена на заседании методической комиссии от 13.12.2023, протокол № 3.