

MINISTRY OF SCIENCE AND HIGHER EDUCATION OF THE RUSSIAN FEDERATION

**Federal State Autonomous Educational Institution of Higher Education
«National Research Lobachevsky State University of Nizhny Novgorod»**

Институт информационных технологий, математики и механики

УТВЕРЖДЕНО

решением президиума Ученого совета ННГУ

протокол № 1 от 16.01.2024 г.

Working programme of the discipline

Natural language processing

Higher education level

Master degree

Area of study / speciality

02.04.02 - Fundamental Informatics and Information Technology

Focus /specialization of the study programme

Artificial Intelligence and Data Analysis

Mode of study

full-time

Nizhny Novgorod

Year of commencement of studies 2024

1. Место дисциплины в структуре ОПОП

Дисциплина Б1.В.ДВ.02.01 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-8: Способен к разработке новых алгоритмических, методических и технологических решений в конкретной сфере профессиональной деятельности	<p>ПК-8.1: Знает методику разработки новых алгоритмических, методических и технологических решений</p> <p>ПК-8.2: Умеет применять полученные знания для разработки новых алгоритмических, методических и технологических решений</p> <p>ПК-8.3: Имеет практический опыт составления технического задания на разработку информационной системы</p>	<p>ПК-8.1: Знать постановки задач автоматической обработки текстов. Знать основные особенности обработки неструктурированных текстов на естественных языках и принципы их анализа на всех уровнях стека лингвистических технологий; основные математические модели и алгоритмы для анализа текста на естественном языке.</p> <p>ПК-8.2: Уметь работать с современными лингвистическими ресурсами (корпусами OpenCorpora, размеченными корпусами ГИКРЯ, семантическим корпусом и т.д.). Уметь использовать методы решения задач автоматической обработки текстов.</p> <p>ПК-8.3: Владеть навыком создания компьютерных программ с использованием современных библиотек, целью которых является решение задач анализа текстов на естественном языке.</p>	Задания	<p>Зачёт:</p> <p>Задания</p> <p>Контрольные вопросы</p>

		Владеть опытом создания компьютерных программ для решения задач автоматической обработки текстов.		
--	--	---	--	--

3. Структура и содержание дисциплины

3.1 Трудоемкость дисциплины

	очная
Общая трудоемкость, з.е.	5
Часов по учебному плану	180
в том числе	
аудиторные занятия (контактная работа):	
- занятия лекционного типа	16
- занятия семинарского типа (практические занятия / лабораторные работы)	16
- КСР	1
самостоятельная работа	147
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			
		Контактная работа (работа во взаимодействии с преподавателем), часы из них			Самостоятельная работа обучающегося, часы
		Занятия лекционного типа	Занятия семинарского типа (практические занятия/лабораторные работы), часы	Всего	
	о ф о	о ф о	о ф о	о ф о	о ф о
Основные задачи и методы	43	4	4	8	35
Синтаксический анализ в естественном языке	43	4	4	8	35
Контекстно-свободные грамматики	43	4	4	8	35
Семантический анализ	50	4	4	8	42
Аттестация	0				
КСР	1			1	
Итого	180	16	16	33	147

--	--	--	--	--	--

Contents of sections and topics of the discipline

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, PyMorphy, 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. Assessment tools for ongoing monitoring of learning progress and interim certification in the discipline (module)

5.1 Model assignments required for assessment of learning outcomes during the ongoing monitoring of learning progress with the criteria for their assessment:

5.1.1 Model assignments (assessment tool - Assignments) to assess the development of the competency ПК-8:

1. 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.
2. Применить алгоритм SKY для получения дерева синтаксического разбора по заданному предложению и грамматике.

Пример

Задана формальная грамматика :

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

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.

Assessment criteria (assessment tool — Assignments)

Grade	Assessment criteria
pass	Работа выполнена в полном объеме и в срок, результаты работы алгоритма корректные на тестовых примерах, результаты работы представлены преподавателю.
fail	Работа не выполнена или выполнена не в полном объеме (программа работает некорректно на тестовых примерах, результаты работы не представлены преподавателю).

5.2. Description of scales for assessing learning outcomes in the discipline during interim certification

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 вследствие отказа обучающегося от ответа	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 несущественных ошибок	Уровень знаний в объеме, соответствующем программе подготовки. Ошибок нет.	Уровень знаний в объеме, превышающем программу подготовки.
<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	Продемонстрированы все основные умения. Решены все основные задачи с негрубыми ошибками. Выполнены все задания в полном объеме, но	Продемонстрированы все основные умения. Решены все основные задачи. Выполнены все задания в полном объеме, но некоторые с	Продемонстрированы все основные умения. Решены все основные задачи с отдельными несущественными недочетами	Продемонстрированы все основные умения. Решены все основные задачи. Выполнены все задания, в полном объеме без недочетов

			объеме	некоторые с недочетами	недочетами	и, выполнены все задания в полном объеме	
<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

Scale of assessment for interim certification

Grade		Assessment criteria
pass	outstanding	All the competencies (parts of competencies) to be developed within the discipline have been developed at a level no lower than "outstanding", the knowledge and skills for the relevant competencies have been demonstrated at a level higher than the one set out in the programme.
	excellent	All the competencies (parts of competencies) to be developed within the discipline have been developed at a level no lower than "excellent",
	very good	All the competencies (parts of competencies) to be developed within the discipline have been developed at a level no lower than "very good",
	good	All the competencies (parts of competencies) to be developed within the discipline have been developed at a level no lower than "good",
	satisfactory	All the competencies (parts of competencies) to be developed within the discipline have been developed at a level no lower than "satisfactory", with at least one competency developed at the "satisfactory" level.
fail	unsatisfactory	At least one competency has been developed at the "unsatisfactory" level.
	poor	At least one competency has been developed at the "poor" level.

5.3 Model control assignments or other materials required to assess learning outcomes during the interim certification with the criteria for their assessment:

5.3.1 Model assignments (assessment tool - Assignments) to assess the development of the competency ПК-8

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.

Assessment criteria (assessment tool — Assignments)

Grade	Assessment criteria
pass	Работа выполнена в полном объеме и в срок, результаты работы алгоритма корректные на тестовых примерах, результаты работы представлены преподавателю.
fail	Работа не выполнена или выполнена не в полном объеме (программа работает некорректно на тестовых примерах, результаты работы не представлены преподавателю).

5.3.2 Model assignments (assessment tool - Control questions) to assess the development of the competency ПК-8

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.

Assessment criteria (assessment tool — Control questions)

Grade	Assessment criteria
pass	Владение основным и дополнительным материалом достаточное или с незначительными ошибками и погрешностями
fail	владение материалом, необходимым по данному предмету, недостаточно. Работу за время 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 02.04.02 - Fundamental Informatics and Information Technology.

Author(s): 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.