

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

**федеральное государственное автономное
образовательное учреждение высшего образования_
«Национальный исследовательский Нижегородский государственный университет
им. Н.И. Лобачевского»**

Высшая школа искусств и дизайна

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

54.04.01 - Дизайн

Направленность образовательной программы

Медиаарт и искусственный интеллект

Форма обучения

очная

г. Нижний Новгород

2024 год начала подготовки

1. Место дисциплины в структуре ОПОП

Дисциплина Б1.В.ДВ.03.02.04 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-5: Готовность демонстрировать наличие комплекса информационно-технологических знаний, владений приемами компьютерного мышления	ПК-5.1: Применяет современные проектные технологии для решения профессиональных задач	ПК-5.1: Знать постановки задач автоматической обработки текстов; 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. Уметь работать с современными лингвистическими ресурсами (корпусами OpenCorpora, размеченными корпусами ГИКРЯ, семантическим корпусом и т.д.); использовать методы решения задач автоматической обработки текстов. Владеть опытом практического использования методов решения задач автоматической обработки текстов	Тест Практическое задание	Зачёт с оценкой: Контрольные вопросы

3. Структура и содержание дисциплины

3.1 Трудоемкость дисциплины

	очная
Общая трудоемкость, з.е.	2
Часов по учебному плану	72
в том числе	
аудиторные занятия (контактная работа):	
- занятия лекционного типа	12
- занятия семинарского типа (практические занятия / лабораторные работы)	22
- КСР	1
самостоятельная работа	37
Промежуточная аттестация	0
	Зачёт с оценкой

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	
	0 ф 0	0 ф 0	0 ф 0	0 ф 0	0 ф 0
Тема 1. Введение в предмет. Основные задачи и методы	5	1	2	3	2
Тема 2. Компьютерная морфология	6	1	2	3	3
Тема 3. Языковая модель	7	1	2	3	4
Тема 4. Исправление опечаток	7	1	2	3	4
Тема 5. Синтаксический анализ в естественном языке	7	1	2	3	4
Тема 6. Грамматика зависимостей	9	1	4	5	4
Тема 7. Контекстно-свободные грамматики (КС-грамматики)	7	1	2	3	4
Тема 8. Статистические методы синтаксического анализа	7	1	2	3	4
Тема 9. Семантический анализ	8	2	2	4	4
Тема 10. Дистрибутивная семантика	8	2	2	4	4
Аттестация	0				
КСР	1				1
Итого	72	12	22	35	37

Содержание разделов и тем дисциплины

Тема 1. 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.

Тема 2. Компьютерная морфология

Морфологический анализ. Словарный и предиктивный морфологический анализ. Лексическая неоднозначность. Инструменты для морфологического анализа и методика их работы (АОТ, PyMorphy, MyStem, NLTK).

Тема 3. 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.

Тема 4. Исправление опечаток

Расстояние Левенштейна, расстояние Левенштейна–Дамерау. Подсчет расстояний Левенштейна. Инструментарий для исправления опечаток. Морфологическая классификация естественных языков. Лингвистическая типология.

Тема 5. Синтаксический анализ в естественном языке

Синтаксическая неоднозначность. Подходы к описанию синтаксиса в естественном языке. Иерархия Хомского. Задача синтаксического разбора.

Тема 6. Грамматика зависимостей

Методы и алгоритмы синтаксического разбора в контексте грамматики зависимостей. Возможности и ограничения грамматики зависимостей.

Тема 7. 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.

Тема 8. Статистические методы синтаксического анализа

Оценка точности синтаксического анализа. Понятие проективности. SyntaxNet.

Тема 9. 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.

Тема 10. Дистрибутивная семантика

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.

Самостоятельная работа, наряду с лекционным курсом и практическими занятиями, является неотъемлемой частью изучения курса. Приступая к изучению дисциплины, студенты должны ознакомиться с учебной программой, учебной, научной и методической литературой, имеющейся в библиотеке, получить в библиотеке рекомендованные учебники и учебно-методические пособия, завести тетради для конспектирования лекций и практических занятий. В ходе самостоятельной работы изучить основную литературу, ознакомиться с дополнительной литературой, научные статьи и материалы социологических исследований. При этом учесть рекомендации преподавателя и требования учебной программы. Подготовить тезисы для выступлений по всем учебным вопросам, выносимым на семинар. Готовясь к докладу или реферативному сообщению, обращаться за методической помощью к преподавателю. При подготовке к зачету повторять пройденный материал в строгом соответствии с учебной программой, примерным перечнем учебных вопросов, выносящихся на экзамен и содержащихся в данной программе. Использовать конспект лекций и литературу, рекомендованную преподавателем. Обратит особое внимание на темы учебных занятий, пропущенных студентом по разным причинам. При необходимости обратиться за консультацией и методической помощью к преподавателю.

В процесс освоения дисциплины выделяют два вида самостоятельной работы:

- аудиторная;
- внеаудиторная.

Аудиторная самостоятельная работа по дисциплине выполняется на учебных занятиях под непосредственным руководством преподавателя и по его заданию.

Внеаудиторная самостоятельная работа выполняется студентом по заданию преподавателя, но без его непосредственного участия. Содержание внеаудиторной самостоятельной работы определяется в соответствии с рекомендуемыми видами заданий согласно рабочей программе учебной дисциплины.

Видами заданий для внеаудиторной самостоятельной работы являются:

- для овладения знаниями: чтение текста (учебника, дополнительной литературы), составление плана текста, конспектирование текста, выписки из текста, учебно-исследовательская работа, использование аудио- и видеозаписей, компьютерной техники и Интернета и др.;
- для закрепления и систематизации знаний: работа с конспектом лекции, обработка текста, повторная работа над учебным материалом, (составление плана, составление таблиц для систематизации учебного материала, ответ на контрольные вопросы, заполнение рабочей тетради, аналитическая обработка текста), подготовка мультимедиа сообщений/докладов к выступлению на семинаре, подготовка реферата, тестирование и др.;
- для формирования умений: решение практических ситуаций и заданий, подготовка к деловым играм, решение тестов и т.д.

Самостоятельная работа может осуществляться индивидуально или группами студентов в

зависимости от цели, объема, конкретной тематики самостоятельной работы, уровня сложности, уровня умений студентов.

Контроль результатов внеаудиторной самостоятельной работы студентов может осуществляться в пределах времени, отведенного на обязательные учебные занятия по дисциплине и внеаудиторную самостоятельную работу студентов по дисциплине, может проходить в письменной, устной или смешанной форме.

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 - Тест) для оценки сформированности компетенции ПК-5:

1. Алгоритм сопоставления каждому конкретному сообщению строго определённой комбинации символов называется

- 1) матрица вероятностей;
- 2) код;
- 3) метод.

2. Комбинация символов алфавита носит название

- 1) вариативное слово;
- 2) комплексное слово;
- 3) кодовое слово.

3. Процесс преобразования сообщения в комбинацию символов в соответствии с кодом называется

- 1) кодированием;
- 2) детерминацией;
- 3) маркировкой.

Критерии оценивания (оценочное средство - Тест)

Оценка	Критерии оценивания
превосходно	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «превосходно»,

Оценка	Критерии оценивания
	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», 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»
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»

5.1.2 Типовые задания (оценочное средство - Практическое задание) для оценки сформированности компетенции ПК-5:

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.

Задание 2

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.

Задание 3

Подсчитать расстояния Левенштейна и Левенштейна-Дамерау для заданных строк, например, Дракон съел собаку. Собака подавилась драконом.

Критерии оценивания (оценочное средство - Практическое задание)

Оценка	Критерии оценивания
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», 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»
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. 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 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	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 и 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 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 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 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 - Контрольные вопросы) для оценки сформированности компетенции ПК-5

Сложность АОТ. Неоднозначность при обработке естественного языка. Уровни неоднозначности.
Основные задачи АОТ
Предмет компьютерной морфологии. Морфологический анализ. Словарный и предиктивный морфологический анализ.
Подходы к определению грамматического значения несловарных слов. Лексическая неоднозначность в морфологическом анализе.
Морфологический анализ на базе правил. Инструменты для морфологического анализа (АОТ, PyMorphy, MyStem) и методика их работы.
Задача частеречевой разметки. Статистическая частеречевая разметка.
Понятие скрытой Марковской модели (НММ). Алгоритм Витерби. Использование алгоритма Витерби для решения задачи частеречевой разметки. Учет незнакомых слов при статистическом подходе к чатеречевой разметке.
Исправление опечаток. Расстояние Левенштейна, расстояние Левенштейна-Дамерау. Подсчет расстояний Левенштейна. Инструментарий.
Морфологическая классификация языков. Примеры.
Синтаксический анализ в естественном языке. Проблематика. Синтаксическая неоднозначность. Подходы

к описанию синтаксиса естественного языка. Иерархия Хомского.
Грамматика зависимостей. Методы. Проблемы (придаточные предложения, и т.д.). Недостаточность ГЗ. Понятие грамматики непосредственно составляющих. Алгоритмы парсинга грамматики НС.
Грамматика непосредственно составляющих. Алгоритмы. Проблема неоднозначности и комбинаторного взрыва.
Алгоритмы статистического парсинга. КС-грамматики. Вероятностные КС-грамматики. Алгоритм СКУ. Оценка качества синтаксического разбора.
Лексикализация. Dependency Parsing. Проективность и непроективность при парсинге. Оценка качества синтаксического разбора ГЗ. SyntaxNet.
Семантический анализ. Модели представления знаний в компьютерной семантике (сетевые модели, концептуальные графы, фреймы и сценарии, современные подходы).
Понятие формальной онтологии. Онтологические ресурсы.
Компьютерные тезаурусы. WordNet, FrameNet. Тезаурусы для русского языка.
Дистрибутивная семантика. Понятие дистрибутивной семантики. Классические count-based подходы к дистрибутивной семантике. Векторное представление слова.
Предиктивные подходы в дистрибутивной семантике. Word2vec. Алгоритмы CBOW и skip-gram. Deep learning и word2vec.
Word2vec. Подход Миколова к ускорению Word2Vec (Hierarchical SoftMax и Negative Sampling). Лингвистические особенности и инструментарий.

Критерии оценивания (оценочное средство - Контрольные вопросы)

Оценка	Критерии оценивания
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», 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

Оценка	Критерии оценивания
	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»
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»

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

Основная литература:

1. Онтологии и тезаурусы: модели, инструменты, приложения / Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. - Москва : ИНТУИТ, 2016., <https://e-lib.unn.ru/MegaPro/UserEntry?Action=FindDocs&ids=663005&idb=0>.

Дополнительная литература:

1. Интеллектуальные робототехнические системы / Афонин В.Л., Макушкин В.А. - Москва : ИНТУИТ, 2016., <https://e-lib.unn.ru/MegaPro/UserEntry?Action=FindDocs&ids=662908&idb=0>.

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

- 1. Python 3.4 или R
- 2. Библиотеки: scikit-learn, NLTK, gensim, tensorflow.
- 3. 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, компьютерами.

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 ФГОС ВО по направлению подготовки/специальности 54.04.01 - Дизайн.

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 26.10.2023, протокол № 6.