

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
президиумом 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
02.04.02 «Фундаментальная информатика и информационные технологии»
(указывается код и наименование направления подготовки / специальности)

Направленность образовательной программы
Искусственный интеллект
(указывается профиль / магистерская программа / специализация)

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

Нижний Новгород

2023 год

1. Место дисциплины в структуре ОПОП

Дисциплина Б1.В.ДВ.06.01 «Обработка естественных языков» относится к дисциплинам по выбору части Блока 1, формируемой участниками образовательных отношений, «Дисциплины (модули)» направления подготовки 02.04.02 «Фундаментальная информатика и информационные технологии», направленность «Искусственный интеллект». Дисциплина преподается в 3 семестре.

№ Варианта	Место дисциплины в учебном плане образовательной программы	Стандартный текст для автоматического заполнения в конструкторе РПД
1	Блок 1. Дисциплины (модули) Часть, формируемая участниками образовательных отношений. Дисциплина по выбору	Дисциплина Б1.В.ДВ.06.01 «Обработка естественных языков» относится к части ООП направления подготовки 02.04.02 Фундаментальная информатика и информационные технологии, формируемой участниками образовательных отношений.

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**	
ПК-8. Способен к разработке новых алгоритмических, методических и технологических решений в конкретной сфере профессиональной деятельности.	ПК-8.1. Знать основы разработки новых алгоритмических, методических и технологических решений в конкретной сфере профессиональной деятельности	<u>Знать</u> постановки задач автоматической обработки текстов. <u>Знать</u> 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. <u>Уметь</u> работать с современными лингвистическими ресурсами (корпусами OpenCorpora, размеченными корпусами ГИКРЯ, семантическим корпусом и т.д.). <u>Уметь</u> использовать методы решения задач автоматической обработки текстов. <u>Владеть</u> опытом практического использования методов решения задач автоматической обработки текстов.	<i>Собеседование</i>

	<p>ПК-8.2. Иметь навыки разработки новых алгоритмических, методических и технологических решений в конкретной сфере профессиональной деятельности</p>	<p><u>Знать</u> основные алгоритмы и методы автоматической обработки текстов. <u>Знать</u> 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. <u>Владеть</u> навыком распознавания возможностей и ограничений существующих на данный момент методов автоматической обработки текстов. <u>Владеть</u> навыками использования программного обеспечения для решения практических задач автоматической обработки текстов.</p>	<p><i>Задачи (практические задания)</i></p>
	<p>ПК-8.3. Иметь навыки управления разработкой и развитием новых алгоритмических, методических и технологических решений в конкретной сфере профессиональной деятельности</p>	<p><u>Знать</u> основные алгоритмы и методы автоматической обработки текстов. <u>Знать</u> 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. <u>Владеть</u> навыком создания компьютерных программ с использованием современных библиотек, целью которых является решение задач анализа текстов на естественном языке. <u>Владеть</u> опытом создания компьютерных программ для решения задач автоматической обработки текстов.</p>	<p><i>Задачи (практические задания)</i></p>

3. Структура и содержание дисциплины

3.1. Трудоемкость дисциплины

	Очная форма обучения
Общая трудоемкость	3 ЗЕТ
Часов по учебному плану	108
в том числе	
аудиторные занятия (контактная работа):	33
- занятия лекционного типа	16
- занятия семинарского типа	16
- занятия лабораторного типа	-
- текущий контроль (КСР)	1
самостоятельная работа	75
Промежуточная аттестация –зачет	

3.2. Содержание дисциплины

Наименование и краткое содержание разделов и тем дисциплины	Всего (часы) Очная	В том числе				
		Контактная работа (работа во взаимодействии с преподавателем), часы. Из них				Самостоятельная работа обучающегося, часы Очная
		Занятия лекционного типа Очная	Занятия семинарского типа Очная	Занятия лабораторного типа Очная	Всего Очная	
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 алгоритм Витерби.	14	2	2		4	10
Исправление опечаток. Расстояние Левенштейна, расстояние Левенштейна–Дамерау. Подсчет расстояний Левенштейна. Инструментарий для исправления опечаток. Морфологическая классификация	14	2	2		4	10

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.	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.	15	2	2		4	11
Текущий контроль (КСР)	1				1	
Промежуточная аттестация – зачет						
Итого	108	16	16		33	75

Практические занятия (лабораторные занятия) 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: выполнение преобразования формата CoNLL-u; разработка PoS-теггера на базе скрытой Марковской цепи и алгоритма Витерби; обучение синтаксического анализатора SyntaxNet; применение модели word2vec для анализа близости семантики двух слов; подсчет расстояния Левенштейна и Левенштейна-Дамерау для заданных строк.

На проведение практических занятий (лабораторных работ) в форме практической подготовки отводится 16 часов.

Практическая подготовка направлена на формирование и развитие:

- практических навыков в соответствии с профилем ОП: Разработка, тестирование, оптимизация программного обеспечения (ПО). Разработка технической документации на продукцию в сфере ИТ.
- компетенций – ПК-8: Способен к разработке новых алгоритмических, методических и технологических решений в конкретной сфере профессиональной деятельности (ПК-8.3: Имеет практический опыт анализа и интерпретации сложных информационных систем).

Текущий контроль успеваемости реализуется в формах опросов на занятиях семинарского типа. Промежуточная аттестация проходит в традиционных формах (экзамен).

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 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 незначительных ошибок	Уровень знаний в объеме, соответствующем программе подготовки, без ошибок.	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 незначительными 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-8

Вопросы	Код формируемой компетенции
1. Сложность АОТ. Неоднозначность при обработке естественного языка. Уровни неоднозначности.	ПК-8
2. Основные задачи АОТ	ПК-8
3. Предмет компьютерной морфологии. Морфологический анализ. Словарный и предиктивный морфологический анализ.	ПК-8
4. Подходы к определению грамматического значения несловарных слов. Лексическая неоднозначность в морфологическом анализе.	ПК-8
5. Морфологический анализ на базе правил. Инструменты для морфологического анализа (АОТ, PyMorph, MyStem) и методика их работы.	ПК-8
6. Задача частеречевой разметки. Статистическая частеречевая разметка.	ПК-8
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-8
8. Исправление опечаток. Расстояние Левенштейна, расстояние Левенштейна-Дамерау. Подсчет расстояний Левенштейна. Инструментарий.	ПК-8
9. Морфологическая классификация языков. Примеры.	ПК-8
10. Синтаксический анализ в естественном языке. Проблематика. Синтаксическая неоднозначность. Подходы к описанию синтаксиса	ПК-8

естественного языка. Иерархия Хомского.	
11. Грамматика зависимостей. Методы. Проблемы (придаточные предложения, и т.д.). Недостаточность ГЗ. Понятие грамматики непосредственно составляющих. Алгоритмы парсинга грамматики НС.	ПК-8
12. Грамматика непосредственно составляющих. Алгоритмы. Проблема неоднозначности и комбинаторного взрыва.	ПК-8
13. Алгоритмы статистического парсинга. КС-грамматики. Вероятностные КС-грамматики. Алгоритм СКУ. Оценка качества синтаксического разбора.	ПК-8
14. Лексикализация. Dependency Parsing. Проективность и непроективность при парсинге. Оценка качества синтаксического разбора ГЗ. SyntaxNet.	ПК-8
15. Семантический анализ. Модели представления знаний в компьютерной семантике (сетевые модели, концептуальные графы, фреймы и сценарии, современные подходы).	ПК-8
16. Понятие формальной онтологии. Онтологические ресурсы.	ПК-8
17. Компьютерные тезаурусы. WordNet, FrameNet. Тезаурусы для русского языка.	ПК-8
18. Дистрибутивная семантика. Понятие дистрибутивной семантики. Классические count-based подходы к дистрибутивной семантике. Векторное представление слова.	ПК-8
19. Предиктивные подходы в дистрибутивной семантике. Word2vec. Алгоритмы CBOW и skip-gram. Deep learning и word2vec.	ПК-8
20. Word2vec. Подход Миколова к ускорению Word2Vec (Hierarchical SoftMax и Negative Sampling). Лингвистические особенности и инструментарий.	ПК-8

5.2.2. Типовые задания/задачи для оценки сформированности компетенции ПК-8

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 СКУ для получения дерева синтаксического разбора по</p>	ПК-8

<p>заданному предложению и грамматике.</p> <p>Пример</p> <p>Задана формальная грамматика :</p> <p>S → NP VP Det → that this the a</p> <p>S → Aux NP VP Noun → book flight meal money tickets</p> <p>S → VP Verb → book include prefer</p> <p>NP → Pronoun Pronoun → I she me</p> <p>NP → Proper-Noun Proper-Noun → Houston NWA</p> <p>NP → Det Nominal Aux → does</p> <p>Nominal → Noun Preposition → from to on near through</p> <p>Nominal → Nominal Noun</p> <p>Nominal → Nominal PP</p> <p>VP → Verb</p> <p>VP → Verb NP</p> <p>VP → Verb NP PP</p> <p>VP → Verb PP</p> <p>VP → VP PP</p> <p>PP → Preposition NP</p> <p>Привести к нормальной форме Хомского и применить алгоритм СКУ для построения дерева составляющих для строки</p> <p>I book the tickets to the Houston.</p> <p><u>Задача 5.</u></p> <p>Подсчитать расстояния Левенштейна и Левенштейна-Дамерау для заданных строк, например,</p> <p>Кошка съела мышку.</p> <p>Мышка съела кошку.</p>	
<p><u>Задание 6</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 7.</u></p> <p>Применить модель word2vec для анализа близости семантики двух слов. Обучить свою модель Skip-gram.</p> <p><u>Задание 8.</u></p> <p>Используя WordNet применить онтологическую модель для анализа семантики текста.</p> <p><u>Задание 9.</u></p> <p>Подсчитать расстояния Левенштейна и Левенштейна-Дамерау для заданных строк, например,</p> <p>Дракон съел собаку.</p> <p>Собака подавилась драконом.</p>	<p>ПК-8</p>

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 ННГУ по направлению 02.04.02 «Фундаментальная информатика и информационные технологии».

Автор д.ф.-м.н., проф. Н. Ю. Золотых

Заведующий кафедрой АГиДМ Н. Ю. Золотых

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