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
01.04.02 Прикладная математика и информатика

Направленность образовательной программы
Компьютерные науки и приложения
(указывается профиль / магистерская программа / специализация)

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

Нижний Новгород

2022 год

Место дисциплины в структуре ОПОП

Дисциплина Б1.В.ДВ.06.02 «Обработка естественных языков» относится к дисциплинам по выбору части Блока 1, формируемой участниками образовательных отношений, «Дисциплины (модули)» направления подготовки «Прикладная математика и информатика», направленность образовательной программы «Компьютерные науки и приложения». Дисциплина преподается в 3 семестре. Трудоемкость дисциплины составляет 3 зачетные единицы, 108 час., зачет.

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

2. 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-4. Способен разрабатывать и анализировать концептуальные и теоретические модели решаемых научных проблем и задач	ПК-4.1. Знает методы разработки и анализа концептуальных и теоретических моделей решаемых научных проблем и задач.	Знать типовые методы применения обработки естественных языков при разработке и анализе концептуальных и теоретических моделей решаемых научных проблем и задач. Уметь применять типовые методы обработки естественных языков при разработке и анализе концептуальных и теоретических моделей решаемых научных проблем и задач.	Задание
	ПК-4.2. Умеет применять методы разработки и анализа концептуальных и теоретических моделей решаемых научных проблем и задач.	Уметь работать с современными лингвистическими ресурсами (корпусами OpenCorpora, размеченными корпусами ГИКРЯ, семантическим корпусом и т.д.). Владеть математическим аппаратом и знаниями об основных структурах данных для решения задач обработки естественных языков.	Задание
	ПК-4.3. Имеет навыки применения методов разработки и анализа концептуальных и	Уметь применять типовые математические методы разработки и анализа концептуальных и теоретических моделей	Задание

	теоретических моделей решаемых научных проблем и задач.	решаемых задач обработки естественных языков. Владеть практическим опытом разработки и применения системного и прикладного программного обеспечения для решения задач обработки естественных языков.	
ПК-11. Способность разрабатывать и анализировать концептуальные и теоретические модели решаемых задач производственно-технологической деятельности	ПК-11.1. Знать методы разработки и анализа концептуальных и теоретических моделей решаемых производственно-технологических задач.	Знать методы разработки и анализа концептуальных и теоретических моделей решаемых задач обработки естественных языков. Владеть навыком анализа концептуальных и теоретических моделей решаемых основных задач автоматической обработки текстов.	Задание
	ПК-11.2. Уметь применять методы разработки и анализа концептуальных и теоретических моделей решаемых производственно-технологических задач.	Уметь решать основные задачи теории систем линейных неравенств и машинного обучения и применять методы разработки и анализа концептуальных и теоретических моделей. Владеть способностью разрабатывать и применять математические методы, системное и прикладное программное обеспечение для решения задач автоматической обработки текстов.	Задание
	ПК-11.3 Иметь навыки применения методов разработки и анализа концептуальных и теоретических моделей решаемых производственно-технологических задач	Уметь применять типовые математические методы и методологии разработки системного и прикладного программного обеспечения для решения задач обработки естественных языков. Владеть способностями осуществлять научное руководство коллективом специалистов, создающих алгоритмы и их программные реализации решения задач автоматической обработки текстов.	Задание

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): **- занятия лекционного типа** **- занятия семинарского типа** **- занятия лабораторного типа** **- текущий контроль (КСР)**	 **16** **–** **16** **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, *n*-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 – экзамен						
Итого	**108**	**16**		**16**	**33**	**75**

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				
Знания	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 ошибки.	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, без ошибок.	Уровень знаний в объеме, превышающем программу подготовки.
Умения	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 отдельными несущест-венным недочетами, выполнены все задания в полном объеме.	Продемонстрированы все основные умения, решены все основные задачи. Выполнены все задания, в полном объеме без недочетов
Навыки	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.	Продемонст-рированы базовые навыки при решении стандартных задач с некоторыми 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 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-4)

Вопросы	*Код формируемой компетенции*
1. Сложность АОТ. Неоднозначность при обработке естественного языка. Уровни неоднозначности.	ПК-4
2. Основные задачи АОТ	ПК-4
3. Предмет компьютерной морфологии. Морфологический анализ. Словарный и предиктивный морфологический анализ.	ПК-4
4. Подходы к определению грамматического значения несловарных слов. Лексическая неоднозначность в морфологическом анализе.	ПК-4
5. Морфологический анализ на базе правил. Инструменты для морфологического анализа (АОТ, PyMorphy, MyStem) и методика их работы.	ПК-4
6. Задача частеречевой разметки. Статистическая частеречевая разметка.	ПК-4
7. Понятие скрытой Марковской модели (HMM). Алгоритм Витерби. Использование алгоритма Витерби для решения задачи частеречевой разметки. Учет незнакомых слов при статистическом подходе к чатеречевой разметке.	ПК-4
8. Исправление опечаток. Рассстояние Левенштейна, расстояние Левенштейна-Дамерау. Подсчет расстояний Левенштейна. Инструментарий.	ПК-4
9. Морфологическая классификация языков. Примеры.	ПК-4
10. Синтаксический анализ в естественном языке. Проблематика. Синтаксическая неоднозначность. Подходы к описанию синтаксиса естественного языка. Иерархия Хомского.	ПК-4
11. Грамматика зависимостей. Методы. Проблемы (придаточные предложения, и т.д.). Недостаточность ГЗ. Понятие грамматики	ПК-4

непосредственно составляющих. Алгоритмы парсинга грамматики НС.	
12. Грамматика непосредственно составляющих. Алгоритмы. Проблема неоднозначности и комбинаторного взрыва.	ПК-4
13. Алгоритмы статистического парсинга. КС-грамматики. Вероятностные КС-грамматики. Алгоритм CKY. Оценка качества синтаксического разбора.	ПК-4
14. Лексикализация. Dependency Parsing. Проективность и непроективность при парсинге. Оценка качества синтаксического разбора ГЗ. SyntaxNet.	ПК-4
15. Семантический анализ. Модели представления знаний в компьютерной семантике (сетевые модели, концептуальные графы, фреймы и сценарии, современные подходы).	ПК-4
16. Понятие формальной онтологии. Онтологические ресурсы.	ПК-4
17. Компьютерные тезаурусы. WordNet, FrameNet. Тезаурусы для русского языка.	ПК-4
18. Дистрибутивная семантика. Понятие дистрибутивной семантики. Классические count-based подходы к дистрибутивной семантике. Векторное представление слова.	ПК-4
19. Предиктивные подходы в дистрибутивной семантике. Word2vec. Алгоритмы CBOW и skip-gram. Deep learning и word2vec.	ПК-4
20. Word2vec. Подход Миколова к ускорению Word2Vec (Hierarchical SoftMax и Negative Sampling). Лингвистические особенности и инструментарий.	ПК-4

5.2.2. Типовые задания/задачи для оценки сформированности компетенции ПК-11

Задания	Компетенция
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 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. Задание 3. Используя WordNet применить онтологическую модель для анализа семантики текста. Задание 4. Применить алгоритм CKY для получения дерева синтаксического разбора по заданному предложению и грамматике.	ПК-11

Пример Задана формальная грамматика : S → NP VP S → Aux NP VP S → VP NP → Pronoun NP → Proper-Noun NP → Det Nominal Nominal → Noun Nominal → Nominal Noun Nominal → Nominal PP VP → Verb VP → Verb NP VP → Verb NP PP VP → Verb PP VP → VP PP PP → Preposition NP Det → that | this | the | a Noun → book | flight | meal | money | tickets Verb → book | include | prefer Pronoun → I | she | me Proper-Noun → Houston | NWA Aux → does Preposition → from | to | on | near | through Привести к нормальной форме Хомского и применить алгоритм CKY для построения дерева составляющих для строки I book the tickets to the Houston. <u>Задача 5</u>. Подсчитать расстояния Левенштейна и Левенштейна-Дамерау для заданных строк, например, Кошка съела мышку. Мышка съела кошку.	
<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. <u>Задание 7</u>. Применить модель word2vec для анализа близости семантики двух слов. Обучить свою модель Skip-gram. <u>Задание 8</u>. Используя WordNet применить онтологическую модель для анализа семантики текста. <u>Задание 9</u>. Подсчитать расстояния Левенштейна и Левенштейна-Дамерау для заданных строк, например, Дракон съел собаку. Собака подавилась драконом.	ПК-11

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 01.04.02 Прикладная математика и информатика.

Автор д.ф.-м.н., проф. _____ Н. Ю. Золотых

Рецензент _____

Заведующий кафедрой АГиДМ_____ Н. Ю. Золотых

Программа одобрена на заседании методической комиссии института информационных технологий, математики и механики от «30» ноября 2022 г. № 13