Приложение 4

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
Бакалавриат

Направление подготовки / специальность
02.03.02 - Фундаментальная информатика и информационные технологии

Направленность образовательной программы
Инженерия программного обеспечения

Форма обучения
очная

г. Нижний Новгород

2022 год начала подготовки

1. Место дисциплины в структуре ОПОП

Дисциплина Б1.В.ДВ.04.05 Обработка естественных языков относится к дисциплинам по выбору.

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-4: Способен проектировать программное обеспечение	*ПК-4.1: Знает типовые решения, библиотеки программных модулей, шаблоны, классы объектов, используемые при разработке программного обеспечения* *ПК-4.2: Знает методы и средства проектирования программного обеспечения* *ПК-4.3.: Знает методы и средства проектирования баз данных* *ПК-4.4.: Умеет использовать существующие типовые решения и шаблоны проектирования программного обеспечения* *ПК-4.5.: Умеет применять методы и средства проектирования программного обеспечения, структур данных, баз данных*	*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.* *Знать основные алгоритмы и методы 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.* *Уметь работать с современными лингвистическими ресурсами (корпусами OpenCorpora, размеченными корпусами ГИКРЯ, семантическим корпусом и т.д.).* *Уметь использовать методы решения задач автоматической обработки текстов.*	*Практическое задание*	*Зачёт: Контрольные вопросы*

		Владеть опытом практического использования методов решения задач автоматической обработки текстов. *Владеть навыком распознавания возможностей и ограничений существующих на данный момент методов автоматической обработки текстов.* *Владеть навыками использования программного обеспечения для решения практических задач автоматической обработки текстов.* *Владеть навыком создания компьютерных программ с использованием современных библиотек, целью которых является решение задач анализа текстов на естественном языке.* *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.	**2**
Часов по учебному плану	**72**
в том числе	
аудиторные занятия (контактная работа):	
- занятия лекционного типа	**16**
- занятия семинарского типа (практические занятия / лабораторные работы)	**16**
- КСР	**1**
самостоятельная работа	**39**
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	

		Занятия лекционного типа	Занятия семинарского типа (практические занятия/лабораторные работы), часы	Всего	Самостоятельная работа обучающегося, часы
	очно	очно	очно	очно	очно
1.Введение в предмет	8	2	2	4	4
2.Компьютерная морфология. Языковая модель.	8	2	2	4	4
3.Исправление опечаток.	8	2	2	4	4
4.Синтаксический анализ в естественном языке. Грамматика зависимостей.	8	2	2	4	4
5.Контекстно-свободные грамматики (КС-грамматики).	9	2	2	4	5
6.Статистические методы синтаксического анализа.	10	2	2	4	6
7.Семантический анализ.	10	2	2	4	6
8.Дистрибутивная семантика.	10	2	2	4	6
Аттестация					
КСР	1			1	
Итого	72	16	16	33	39

Содержание разделов и тем дисциплины

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.
2. 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.
3. Исправление опечаток. Расстояние Левенштейна, расстояние Левенштейна– Дамерау. Подсчет расстояний Левенштейна. Инструментарий для исправления опечаток. Морфологическая классификация естественных языков. Лингвистическая типология.
4. 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.
5. 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.
6. Статистические методы синтаксического анализа. Оценка точности синтаксического анализа. Понятие проективности. SyntaxNet.
7. 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.
8. 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.

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 - Практическое задание) для оценки сформированности компетенции ПК-4:

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.

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.

Задача 5.

Подсчитать расстояния Левенштейна и Левенштейна-Дамерау для заданных строк, например,

Кошка съела мышку.

Мышка съела кошку.

Задание 6

- 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 7.

Применить модель word2vec для анализа близости семантики двух слов. Обучить свою модель Skip-gram.Задание 8.
Используя WordNet применить онтологическую модель для анализа семантики текста.

Задание 9.

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
зачтено	Выполнены все или большая часть этапов решения задачи или задача решена с незначительными недочетами. Код и результаты работы представлены преподавателю в срок.
не зачтено	Выполнены не все лабораторные работы или выполнены не в полном объеме (представлено не полное описание этапов выполнения заданий, код работает некорректно, 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**				
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 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.

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 несущественными 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 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	Продемонстрированы 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 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-4

1. Сложность АОТ. Неоднозначность при обработке естественного языка. Уровни неоднозначности.
2. Основные задачи АОТ

3. Предмет компьютерной морфологии. Морфологический анализ. Словарный и предиктивный морфологический анализ.
4. Подходы к определению грамматического значения несловарных слов. Лексическая неоднозначность в морфологическом анализе.
5. Морфологический анализ на базе правил. Инструменты для морфологического анализа (АОТ, PyMorphy, MyStem) и методика их работы.
6. Задача частеречевой разметки. Статистическая частеречевая разметка.
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.
8. Исправление опечаток. Рассстояние Левенштейна, расстояние Левенштейна-Дамерау. Подсчет расстояний Левенштейна. Инструментарий.
9. Морфологическая классификация языков. Примеры.
10. Синтаксический анализ в естественном языке. Проблематика. Синтаксическая неоднозначность. Подходы к описанию синтаксиса естественного языка. Иерархия Хомского.
11. Грамматика зависимостей. Методы. Проблемы (придаточные предложения, и т.д.). Недостаточность ГЗ. Понятие грамматики непосредственно составляющих. Алгоритмы парсинга грамматики НС.
12. Грамматика непосредственно составляющих. Алгоритмы. Проблема неоднозначности и комбинаторного взрыва.
13. Алгоритмы статистического парсинга. КС-грамматики. Вероятностные КС-грамматики. Алгоритм CKY. Оценка качества синтаксического разбора.
14. Лексикализация. Dependency Parsing. Проективность и непроективность при парсинге. Оценка качества синтаксического разбора ГЗ. SyntaxNet.
15. Семантический анализ. Модели представления знаний в компьютерной семантике (сетевые модели, концептуальные графы, фреймы и сценарии, современные подходы).
16. Понятие формальной онтологии. Онтологические ресурсы.
17. Компьютерные тезаурусы. WordNet, FrameNet. Тезаурусы для русского языка.
18. Дистрибутивная семантика. Понятие дистрибутивной семантики. Классические count-based подходы к дистрибутивной семантике. Векторное представление слова.
19. Предиктивные подходы в дистрибутивной семантике. Word2vec. Алгоритмы CBOW и skip-gram. Deep learning и word2vec.
20. 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	Студент дает полный ответ на все теоретические вопросы, возможно с незначительными неточностями в определении понятий, процессов и т.п. Студент работал на практических занятиях и выполнил все задания для текущего контроля успеваемости как минимум на 80%.
не зачтено	Студент дает ошибочные ответы, как на теоретические вопросы, так и на наводящие вопросы экзаменатора. Студент пропустил большую часть практических занятий и не выполнил задания для текущего контроля успеваемости.

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):

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, 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 ОС ННГУ по направлению подготовки/специальности 02.04.02 - Фундаментальная информатика и информационные технологии.

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

Рецензент(ы): Старостин Николай Владимирович, доктор технических наук.

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

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