Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Институт прикладной математики и компьютерных наук

УТВЕРЖДАЮ:
Директор
_____ А. В. Замятин
«16» мая 2022 г.

Рабочая программа дисциплины

Обработка естественного языка- I

по направлению подготовки

02.04.02 Фундаментальная информатика и информационные технологии

Направленность (профиль) подготовки :
Моделирование систем искусственного интеллекта

Форма обучения
Очная

Квалификация
Магистр

Год приема
2022

Код дисциплины в учебном плане: ФТД.04

СОГЛАСОВАНО:
Руководитель ОП
_____А.Н. Моисеев

Председатель УМК
_____С.П. Сущенко

Томск – 2022

1. Цель и планируемые результатыосвоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

УК-1 – Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий

УК-3 – Способен организовывать и руководить работой команды, вырабатывая командную стратегию для достижения поставленной цели

ПК-4 – Способен управлять получением, хранением, передачей, обработкой больших данных.

ПК-5. Способен исследовать и разрабатывать архитектуры систем искусственного интеллекта для различных предметных областей на основе комплексов методов инструментальных средств систем искусственного интеллекта.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИУК-1.1. Выявляет проблемную ситуацию, на основе системного подхода осуществляет ее многофакторный анализ и диагностику.

ИУК-1.2. Осуществляет поиск, отбор и систематизацию информации для определения альтернативных вариантов стратегических решений в проблемной ситуации.

ИУК-1.3. Предлагает и обосновывает стратегию действий с учетом ограничений, рисков и возможных последствийИУК-3.1. Формирует стратегию командной работы на основе совместного обсуждения целей и направлений деятельности для их реализации.

ИУК-3.2. Организует работу команды с учетом объективных условий (технология, внешние факторы, ограничения) и индивидуальных возможностей членов команды.

ИУК-3.3. Обеспечивает выполнение поставленных задач на основе мониторинга командной работы и своевременного реагирования на существенные отклоненияИПК-4.1. Осуществляет мониторинг и оценку производительности обработки больших данных.

ИПК-4.2. Использует методы и инструменты получения, хранения, передачи, обработки больших данных.

ИПК-4.3. Разрабатывает предложения по повышению производительности обработки больших данных.

ИПК-5.1. Исследует и разрабатывает архитектуры систем искусственного интеллекта для различных предметных областей.

2. Задачи освоения дисциплины

– рассмотреть классификацию задач обработки естественного языка,

– научить производить сегментацию текста с использованием доступных инструментов.

3. Место дисциплины в структуре образовательной программы

Дисциплина относится к факультативным дисциплинам.

4. Семестр(ы)освоения и форма(ы) промежуточной аттестации по дисциплине

Второй семестр, экзамен

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются компетенции, сформированные в ходе освоения образовательных программ предшествующего уровня образования.

6. Язык реализации

Русский

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 5 з.е., 180 часов, из которых:
-лекции: 20 ч.
-лабораторные: 40 ч.
Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Классификация задач и оценка качества решений.

Введение в компьютерную лингвистику и обработку естественного языка. Классификация задач обработки естественного языка. Оценка качества решений задач обработки естественного языка.

Тема 2. Токенизация и сегментация.

Цели предварительной обработки текста. Текстовый анализ. Токенизация. Сегментация текста.

Тема 3. Лемматизация и формальные грамматики.

Задача лемматизации и инструменты для её решения. Формальные грамматики и Томита-парсер.

Тема 4. Векторное представление слов и Word2vec.

Векторное представление слов. Word2Vec, Glove, FastText.

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится путем проведения контрольных работ, проверки выполнения заданий по лабораторным работам и фиксируется в форме контрольной точки не менее одного раза в семестр.

10. Порядок проведения и критерии оценивания промежуточнойаттестации

Промежуточная аттестация проводится в форме экзамена. Результаты экзамена определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

«Отлично» – студент выполнил все лабораторные работы, нет неудовлетворительных оценок за контрольные работы, средняя (округленная) оценка за контрольные работы – «отлично»;

«Хорошо» – студент выполнил все лабораторные работы, нет неудовлетворительных оценок за контрольные работы, средняя (округленная) оценка за контрольные работы – «хорошо»;

«Удовлетворительно» – студент выполнил все лабораторные работы, нет неудовлетворительных оценок за контрольные работы, средняя (округленная) оценка за контрольные работы – «удовлетворительно»;

«Неудовлетворительно» – студент не сдал лабораторные работы или сдал хотя бы одну контрольную работу на «неудовлетворительно».

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle».

б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине.

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

1. Ханнес Хапке, Коул Ховард, Хобсон Лейн. Обработка естественного языка в действии. – СПб.: Питер, 2020 — 576 с.

2. Ганегедара Т. Обработка естественного языка с TensorFlow / пер. с анг. В. С. Яценкова. – М.: ДМК Пресс, 2020. – 382 с.

б) дополнительная литература:

1. Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python. O'Reilly Media. 2009. 512 с.

в) ресурсы сети Интернет:

Наименование	**Ссылка на ресурс**	**Доступность (свободный доступ/ ограниченный доступ)**
1	***2***	***3***
Информационно-справочные системы		
Обработка естественного языка	https://moodle.ido.tsu.ru/course/view.php?id=1399	Свободный доступ
Основы Natural Language Processing для текста	https://habr.com/ru/company/Voximplant/blog/446738/	Свободный доступ
Обработка естественного языка	https://medium.com/nuances-of-programming/обработка-естественного-языка-b1e1cf606929	Свободный доступ
Электронно-библиотечные системы		
Научная библиотека ТГУ	https://www.lib.tsu.ru/	Свободный доступ
Электронно-библиотечная система «Лань»	https://e.lanbook.com/	Общедоступная с авторизацией, по подписке
КиберЛенинка	https://cyberleninka.ru/	Свободный доступ
Профессиональные базы данных		
Искусственный интеллект и сферы его применения. Новости разработки квантовых компьютеров. Исследования искусственных нейронных сетей.	https://ai-news.ru	Свободный доступ

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:
– Python
– Проект Jupyter
– DeepPavlov
– Проект Natasha
– Yandex DataSphere

б) информационные справочные системы:
– Электронный каталог Научной библиотеки ТГУ – http://chamo.lib.tsu.ru/search/query?locale=ru&theme=system
– Электронная библиотека (репозиторий) ТГУ – http://vital.lib.tsu.ru/vital/access/manager/Index
– ЭБС Лань – http://e.lanbook.com/
– ЭБС Консультант студента – http://www.studentlibrary.ru/
– Образовательная платформаЮрайт – https://urait.ru/

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения лабораторных занятий с установленным необходимым программным обеспечением.

Помещения для самостоятельной работы, оснащенные компьютерной техникой и доступом к сети Интернет, в электронную информационно-образовательную среду и к информационным справочным системам.

15. Информация о разработчиках

Пожидаев Михаил Сергеевич, канд. техн. наук, кафедра теоретических основ информатики НИ ТГУ, доцент