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

Институт прикладной математики и компьютерных наук

УТВЕРЖДЕНО:
Директор
А. В. Замятин

Рабочая программа дисциплины

Интеллектуальный анализ текста

по направлению подготовки

02.04.02 Фундаментальная информатика и информационные технологии

Направленность (профиль) подготовки:
Математика беспроводных сетей связи и интернета вещей

Форма обучения
Очная

Квалификация
Магистр

Год приема
2024

СОГЛАСОВАНО:
Руководитель ОП
С.П. Моисеева

Председатель УМК
С.П. Сущенко

Томск – 2024

1. Цель и планируемые результаты освоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

ОПК-1 Способен находить, формулировать и решать актуальные проблемы прикладной математики, фундаментальной информатики и информационных технологий.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-1.1 Анализирует проблемы в области прикладной математики, фундаментальной информатики и информационных технологий

ИОПК-1.3 Решает актуальные проблемы прикладной математики, фундаментальной информатики и информационных технологий

2. Задачи освоения дисциплины

– Освоить классические методы анализа текста на естественном языке;

– Получить понимание основ векторного представления слов и применение его на практике;

– Получить понимание основ и применения на практике рекуррентной модели долгой и краткосрочной памяти;

– Получить понимание основ трансформернной модели и как применять ее на практике.

3. Место дисциплины в структуре образовательной программы

Дисциплина относится к Блоку 1 «Дисциплины (модули)».

Дисциплина относится к обязательной части образовательной программы. Дисциплина входит в модуль Общепрофессиональные дисциплины.

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

Третий семестр, зачет

5. Входные требования для освоения дисциплины

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

Для успешного освоения дисциплины требуется знание линейной алгебры, методов трансляции, теории вероятности, языка Python.

6. Язык реализации

Русский

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 3 з.е., 108 часов, из которых:
-лекции: 8 ч.
-лабораторные: 20 ч.

Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Введение и классические алгоритмы.
Понятия токенизации, сегментации, лемматизации и стемминга.
Формальные аналитические грамматики и утилита Томита-парсер.
TF-iDF, скрытые марковские модели и алгоритм Витерби.

Тема 2. Линейная ячейка и Word2vec.
Перцептрон, полносвязные сети и функции активации.
Векторное представление слов.
Модель Word2vec.

Тема 3. Рекуррентные ИНС и модели памяти.
Идея рекуррентной сети и её особенности.
Нейронная сеть Элмана.
Модель Seq2seq.

Долгая краткосрочная память.
Тема 4. Механизм внимания.
Идея механизма внимания.
Подходы Богданова и Луонга.
Внутреннее внимание.
multi-head attention и позиционное кодирование.

Тема 5. Трансформер.
Преимущества Трансформера и его назначение.
Схема кодера.
Схема декодера.
Типы внутреннего внимания в модели Трансформера.

Тема 6. BERT и GPT.
Описание модели BERT.
Идея fine tuning.
Описание семейства GPT.
Сравнение BERT и GPT между собой.

Тема 7. Прикладные аспекты использования LLM.
Построения промтов для генеративных моделей.
Голосовые ассистенты.
Тесты и оценка качества решения языковых задач.

Тема 8. Изображение и звук.
Диффузионный процесс и Stable diffusion.
Мел-кепстральные коэффициенты.
Синтез речи и модель Tacotron2.

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится путем проверки выполнения практических работ, и фиксируется в форме контрольной точки не менее одного раза в семестр.

Оценочные материалы текущего контроля размещены на сайте ТГУ в разделе «Информация об образовательной программе» - https://www.tsu.ru/sveden/education/eduop/.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Зачет в третьем семестре проводится в устной форме по билетам. Экзаменационный билет состоит из трех вопросов. Продолжительность зачета 1 час.

Оценочные материалы для проведения промежуточной аттестации размещены на сайте ТГУ в разделе «Информация об образовательной программе» - https://www.tsu.ru/sveden/education/eduop/.

11. Учебно-методическое обеспечение

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

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

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Смирнов, Иван Валентинович. Интеллектуальный анализ текстов на основе методов разноуровневой обработки естественного языка / И. В. Смирнов. — Москва : ФИЦ ИУ РАН, 2023. — 354 с. : ил., табл.; 22 см.; ISBN 978-5-6050647-0-1

б) дополнительная литература:

– Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Łukasz; Polosukhin, Illia. Attention Is All You Need. 2017 https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

в) ресурсы сети Интернет:

– Российская государственная библиотека: https://search.rsl.ru/

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

– Microsoft Office Standart 2013 Russian: пакет программ. Включает приложения: MS Office Word, MS Office Excel, MS Office PowerPoint, MS Office On-eNote, MS Office Publisher, MS Outlook, MS Office Web Apps (Word Excel MS PowerPoint Outlook);

– публично доступные облачные технологии (Google Docs, Яндекс диск и т.п.).

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 платформа Юрайт – https://urait.ru/

– ЭБС ZNANIUM.com – https://znanium.com/

– ЭБС IPRbooks – http://www.iprbookshop.ru/

в) профессиональные базы данных *(при наличии)*:

– Университетская информационная система РОССИЯ – https://uisrussia.msu.ru/

– Единая межведомственная информационно-статистическая система (ЕМИСС) – https://www.fedstat.ru/

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения занятий семинарского типа, индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации.

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, доцент кафедры теоретических основ информатики.