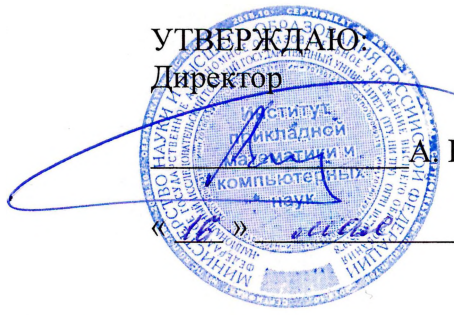
Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Институт прикладной математики и компьютерных наук



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

Рабочая программа дисциплины

Обработка естественного языка

по направлению подготовки

02.03.02 Фундаментальная информатика и информационные технологии

Направленность (профиль) подготовки:
Искусственный интеллект и разработка программных продуктов

Форма обучения
Очная

Квалификация
Бакалавр

Год приема
2022

Код дисциплины в учебном плане: Б1.В.02.05

СОГЛАСОВАНО:
Руководитель ОП
_____А.В. Замятин

Председатель УМК
_____С.П. Сущенко

Томск – 2022

1. Цель и планируемые результаты освоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

– ОПК-2 – способность применять компьютерные/суперкомпьютерные методы, современное программное обеспечение, в том числе отечественного происхождения, для решения задач профессиональной;

– ПК-1 – способность осуществлять программирование, тестирование и опытную эксплуатацию ИС с использованием технологических и функциональных стандартов, современных моделей и методов оценки качества и надежности программных средств.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-2.3. Использует инструментальные средства высокопроизводительных вычислений в научной и практической деятельности.

ИОПК-2.2. Использует методы высокопроизводительных вычислительных технологий, современного программного обеспечения, в том числе отечественного происхождения.

ИОПК-2.1. Обладает необходимыми знаниями основных концепций современных вычислительных систем.

ИПК-1.3. Кодирует на языках программирования и проводит модульное тестирование ИС.

2. Задачи освоения дисциплины

Цель дисциплины - обучить студентов передовым методам, моделям, средствам и технологиям компьютерной обработки текстов на естественных языках дать умение представлять в алгоритмическом виде процессы анали-за и синтеза текста.

Задачи дисциплины:

– получение теоретических знаний и практических навыков обработки естественно-языковых текстов;

– знание сложностей, связанных с применением существующих методов обработки естественно-языковых текстов;

– умение использовать полученные знания по разработке, адаптации и использованию новейших средств для обработки текстов на естественных языках;

– научить студентов проводить аналитическое исследование и разрабатывать приложения с применением технологий обработки естественного языка в соответствии с требованиями заказчика.

3. Место дисциплины в структуре образовательной программы

Дисциплина относится к части образовательной программы, формируемой участниками образовательных отношений, предлагается обучающимся на выбор. Дисциплина входит в модуль Искусственный интеллект.

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

Седьмой семестр, зачет

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются результаты обучения по следующим дисциплинам: «Основы программирования», «Алгоритмы и структуры данных», «Интеллектуальные системы», «Визуализация многомерных данных», «Статистические методы машинного обучения».

6. Язык реализации

Русский

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 3 з.е., 108 часов, из которых:
-лекции: 16 ч.
-практические занятия: 32 ч.
Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Введение, история развития дисциплины, решаемые задачи, подходы, методы и инструменты

Раскрываются три основных этапа развития технологий обработки естественного языка: словарные, вероятностные и интеллектуальные алгоритмы. Даётся классификация задач. Описываются основные методы реализации алгоритмов: локальные, облачные сервисы.

Тема 2. Предварительна обработка текстовых данных

Поясняются назначение и типы предварительной обработки текста: сегментация, токенизация, лемматизация. Сравниваются лемматизация и стемминг. Поясняется роль лемматизации в построении поисковых. индексов. Пояснение недетерминированности проведения сегментации и токенизации.

Тема 3. Вероятностные алгоритмы

Приводятся основные черты вероятностных алгоритмов. Поясняется их роль в современных системах. В качестве примеров приводятся скрытые марковские модели, алгоритм Витерби, EM-алгоритм. Для описания EM-алгоритма поясняется назначение тематического моделирования.

Тема 4. Формальные грамматики

Определение аналитических формальных грамматик по Хомскому. Раскрытие их особенностей и принципиальных ограничений. Примеры задач, которые в настоящий момент можно решать при помощи формальных грамматик. Пояснений функций утилиты Томита-парсер.

Тема 5. Векторное представление слов

Поясняется идея замены слов точками в векторном пространстве. Приводятся примеры алгебраических операций над словами, заменёнными. точками. Определение семантической близости слов через метрики в векторном пространстве. Способы получения векторного представления. Модель Word2vec.

Тема 6. Модель Seq2seq

Пояснение преобразования последовательностей через рекуррентные ячейки. Понятия кодера и декодера. Идея долгой краткосрочной памяти. Идея дополнения кодера и декодера связью через механизм внимания.

Тема 7. Self-attention и Трансформер

Обоснование недостатков модели Seq2seq. Введение понятия Self-attention и пояснение его преимуществ. Назначение ячеек query, key и value. Описание модели Трансформер. Основные преимущества. Описание структуры кодера и декодера Трансформера.

Тема 8. BERT и GPT-3

Описание возможностей построения новых моделей на трансформере. Раздельное использование кодера и декодера. Модель BERT. Идея fine tuning. Модель GPT-3. Применение GPT-3 в практических задачах.

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится путем контроля посещаемости, проведения контрольных работ, тестов по лекционному материалу, и фиксируется в форме контрольной точки не менее одного раза в семестр.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Промежуточная аттестация проводится в форме зачета. Результаты зачета – оценки «зачтено», «не зачтено» проставляется по результатам сдачи практических работ.

Перечень практических работ:

Практическая работа № 1. Парсинг сайтов / использование api для получения текстовых данных.

Практическая работа № 2. Реализация стеммера Портера.

Практическая работа № 3. Использование библиотек для морфологического анализа, решение задачи частеречной разметки.

Практическая работа № 4. Векторное представление текста, word2vec, модели skip-gram и CBOW.

Практическая работа № 5. Тематическое моделирование с использованием библиотеки gensim.

Практическая работа № 6. Анализ тональности текстовых данных. Развертывание обученной модели в вебе.

Практическая работа № 7. Построение языковой модели, порождение текста.

Практическая работа № 8. Генерация подписи к изображению.

Итоговая оценка по предмету выставляется на основе результатов проверки практических работ:

«зачтено» – студент выполнил все практические работы, ответил на все вопросы по практической работе;

«не зачтено» – студент не сдал какие-либо практические работы, не ответил на вопросы по практической работе.

Во время зачета студент может повысить свою оценку, сдав заново соответствующую практическую работу.

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle» - https://moodle.tsu.ru/course/view.php?id=22124

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

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.

– Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. — М.: Изд-во НИУ ВШЭ, 2017. — 269 с.

– Введение в когнитивную лингвистику: учебное пособие. Изд. 2-е, перераб. — Калининград: Изд-во БФУ им. И. Канта, 2012. — 313 с.

– Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. — СПб.: Питер, 2018. — 480 с.: ил. — (Серия «Библиотека программиста»).

– Хобсон Лейн, Ханнес Хапке, Коул Ховард Обработка естественного языка в действии. — СПб.: Питер, 2020. — 576 с.: ил. — (Серия «Для профессионалов»).

– Li Deng Yang Liu Deep Learning in Natural Language Processing. ISBN 978-981-10-5209-5 https://doi.org/10.1007/978-981-10-5209-5

– Николаев И.С., Митренина О.В., Ландо Т.М. Прикладная и КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА. URSS. 2017. 320 с. ISBN 978-5-9710-4633-2.

– Ян Гудфеллоу, Иошуа Бенджио, Аарон Курвилль. Глубокое обучение. Второе цветное издание, исправленное. М.: ДМК Пресс, 2018. – 652 с.

– Франсуа Шолле. Глубокое обучение на Python. СПб: Питер, 2018. – 400 с.

– Daniel Jurafsky, James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, 2008. – 1044 с

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

– Microsoft Visual Studio;

– 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

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения практических занятий, 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.