

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

« 15 » июня 20 23 г.

Рабочая программа дисциплины

Обработка естественного языка - I

по направлению подготовки

01.04.02 Прикладная математика и информатика

Направленность (профиль) подготовки:

Big Data and Data Science

Форма обучения

Очная

Квалификация

Магистр

Год приема

2023

Код дисциплины в учебном плане: Б1.О.06

СОГЛАСОВАНО:

Руководитель ОП

А.В. Замятин

Председатель УМК

С.П. Сущенко

Томск – 2023

1. Цель и планируемые результаты освоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

- УК-1 – способность осуществлять критический анализ проблемных ситуаций на основе системного подхода, выработать стратегию действий;
- УК-3 – способность организовывать и руководить работой команды, выработывая командную стратегию для достижения поставленной цели;
- ПК-6 – способность управлять получением, хранением, передачей, обработкой больших данных.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИУК-1.1 Выявляет проблемную ситуацию, на основе системного подхода осуществляет её 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 последствий.

ИУК-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 отклонения.

ИПК-6.1 Осуществляет мониторинг и оценку производительности обработки больших данных.

ИПК-6.2 Использует методы и инструменты получения, хранения, передачи, обработки больших данных.

ИПК-6.3 Разрабатывает предложения по повышению производительности обработки больших данных.

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 анализа 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 обязательной части образовательной программы.

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

6. Язык реализации

Английский

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 5 з.е., 180 часов, из которых:

-лекции: 20 ч.

-практические занятия: 40 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. Дается классификация задач. Описываются основные методы реализации алгоритмов: локальные, облачные сервисы.

Тема 2. Предварительная обработка текстовых данных

Поясняются назначение и типы предварительной обработки текста: сегментация, токенизация, лемматизация. Сравняются 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. Пояснение 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 экзамена.

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 SBOW.

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:

«Отлично»: учащийся свободно владеет пройденным материалом, отвечает правильно на подавляющее большинство преподавателей, способен самостоятельно развивать решение предложенных проблем и показывает наличие знаний, полученных вне занятий.

«Хорошо»: учащийся владеет необходимым материалом и отвечает на некоторые вопросы преподавателя. Способен решать поставленные задачи с умеренным количеством подсказок.

«Удовлетворительно»: Учащийся слабо владеет материалом и практически не может ответить на вопросы преподавателя. С решением поставленных задач справляется только при наличии интенсивной помощи преподавателя. Знания, полученные вне уроков, отсутствуют.

«Неудовлетворительно»: Какие-либо знания по предмету отсутствуют. При формулировке вопросов и задач не ориентируется в природе решаемой проблемы. На вопросы преподавателей ответить не в состоянии.

Во время экзамена 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., Кадури 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, Йошуа 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.