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

Филологический факультет

УТВЕРЖДЕНО:
Декан
И. В.Тубалова

Рабочая программа дисциплины

Искусственный интеллект в задачах обработки естественного языка

по направлению подготовки

45.03.03 Фундаментальная и прикладная лингвистика

Направленность (профиль) подготовки:
Фундаментальная и прикладная лингвистика

Форма обучения
Очная

Квалификация
Бакалавр

Год приема
2025

СОГЛАСОВАНО:
Руководитель ОП
А.В. Васильева

Председатель УМК
Ю.А. Тихомирова

Томск – 2025

1. Цель и планируемые результаты освоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

ПК-4 Способен разрабатывать программный код при решении задач автоматической обработки текстов.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИПК-4.1 Применяет способы формализации и алгоритмизации поставленных задач в сфере автоматической обработки текстов

ИПК-4.2 Создает программный код с использованием языков программирования и манипулирования данными в сфере автоматической обработки текстов

2. Задачи освоения дисциплины

– Освоить математический аппарат анализа языковых явлений на базе методов искусственного интеллекта.

– Разрабатывать корпус, знать и понимать основные принципы работы корпусной лингвистики

– Научиться разрабатывать и применять на практике методы извлечения фактов из текста

– Освоить базовые принципы работы с текстовым массивом данных

– Разбираться и применять на практике различные типы нейронных сетей в задачах обработки естественного языка

3. Место дисциплины в структуре образовательной программы

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

Дисциплина относится к части образовательной программы, формируемой участниками образовательных отношений, предлагается обучающимся на выбор.

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

Седьмой семестр, экзамен

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 требуются результаты обучения по следующим дисциплинам: лингвистические базы данных, обработка естественного языка, теория вероятностей, машинное обучение

6. Язык реализации

Русский

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 3 з.е., 108 часов, из которых:

-лекции: 12 ч.

-практические занятия: 22 ч.

в том числе практическая подготовка: 22 ч.

Объем самостоятельной работы студента определен учебным планом.

Тема 1. Введение в компьютерную лингвистику: цели, задачи, основные понятия, направления. NLP: современные тенденции и проблемы.

Блок 1. Корпусная лингвистика

Тема 1.1. Введение в корпусную лингвистику. Введение в корпусную лингвистику

Типология корпусов.

Разметка корпуса: лингвистическая и экстралингвистическая разметки

Основные свойства и виды корпусов: НКРЯ, КРУТ, RLC

Тема 1.2. Разработка корпуса: готовые инструменты (tsacorpus, antconc, SctechEngine); проектирование и разработка корпуса: SQL/NoSQL, Docker

Квантитативные методы исследования в корпусной лингвистике: LL-score, IPM, loglikelihood, метрики для определения ключевых слов (TF-IDF), n-грамы.

Тема 1.3. Морфологическая разметка корпуса: mystem, AOT, TreeTagger, Marmot, UDIPIPE

Синтаксическая разметка корпуса: UDIPIPE

Аннотированная разметка корпуса с помощью ELAN

Тема 1.4. Итоговая работа: разработка корпуса (SQL, flask, Docker)

Блок 2. NLP – Обработка естественного языка

Тема 2.1. Парсинг и структуризация текстового массива данных: BeatifulSoup, Selenium.

Тема 2.2. Предобработка текстового массива данных: стеминг/лемматизация, операции над строками. Исправление ошибок (грамматических) в тексте.

Тема 2.3. Векторизация (WordEmbedings) – OneHotEncoding, Bag of Words, word2vec, FastText, Glove, BERT, Transformers, T5

Тема 2.4. Математические операции над векторами. Семантическая близость, снятие омонимии.

Тема 2.4. Задачи классификации: ML, BERT

Тема 2.5. Извлечение данных из текста: контекстно-свободные грамматики (Tomita-parser, Natasha) vs transformers: Spacy

Тема 2.5. Задачи генерации и суммаризации: Seq2Seq, GPT

Тема 2.6. Задачи перевода: Seq2Seq

Тема 2.7. Автоматическое распознавание речи

Тема 2.8. Мультимолдальные системы: Visual Question Answering (VQA)

Тема 2.9. Дообучение моделей (GPT, Whisper). Репозиторий huggingface

Тема 2.10. Итоговая работа: разработка модели NLP

Блок 3. Чат-боты

Тема 3.1 Скриптовые чат-боты

Тема 3.2 Чат-боты на основе

9. Текущий контроль по дисциплине

Текущий контроль образовательной программы (блока, темы, раздела, модуля) требованиям образовательных стандартов по направлениям подготовки/специальностям. Текущий контроль успеваемости обучающихся направлен на определение соответствия результатов обучения после освоения элемента по дисциплине проводится путем контроля посещаемости, проведения контрольных работ, тестов по лекционному материалу, разработки кода, выполнения домашних заданий и фиксируется в форме контрольной точки не менее одного раза в семестр. Примерные задания текущего контроля:

1. Напишите парсер для сайта ria.tomsk.ru. Структурируйте текст в формате csv/json/tsv по следующим атрибутам: ссылка, дата публикации, заголовок, текст, рубрика.

2. Создайте стартеру БД (нотация Чена, нотация Мартина), учитывающую морфологическую разметку текстов. Создайте корпус текстов из п 1.

2. Дообучите модель суммаризации заголовка новости из текста на базе модели ruGPT

3. Напишите контекстно-свободные грамматики для Томиты-парсер, извлекающие данные по конструкту: {ФИО-Должность}-[Галажинский Э.В. – ректор ТГУ].

4. Извлеките из спортивных новостей данные по структуре: {Команда1-Команда-Реультат-Счет} – [ЦСК-«Крылья советов» - победа-2:1]

5. Сравните статистическую меру TF-IDF у двух текстов из НКРЯ, статистически оцените важность слова (лексемы) в контексте документа.

Для повторения метода TF-IDF обратите внимание на презентацию TF-IDF.ppt с главной страницы курса.

Выберите поиск одного слова в любом подкорпусе НКРЯ (или в двух разных подкорпусах). Слово вводите в поисковой форме форме "Лексико-грамматический поиск", а не в поиске точных форм. Никаких дополнительных грамматических ограничений устанавливать не нужно. Сравните важность этого слова в двух разных текстах. Количество словоформ в тексте можно увидеть в перечне параметров текста, если кликнуть по названию текста.

Образцы оформления и вычисления см. в файле IPM_TF-IDF.xlsx

Обязательно приложите и вставьте под соответствующими таблицами скриншоты с метатекстовой информацией (см. в образце) для подтверждения достоверности числовых данных из текстов.

6. Разработайте чат-бот для туристической компании. Бот должен опираться на БД, извлекать дату, геолокацию. Полученные данные формировать в SQL, а ответ пользователю. возвращать – plaint text.

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 с оценкой в восьмом семестре проводится в письменной форме по билетам. Экзаменационный билет состоит из трех частей. Продолжительность зачета с оценкой 1 час. Первая часть представляет собой ответ на теоретический вопрос в устной форме, проверяющую компетенции ПК-3, ПК-4, ИПК-3.1, ИПК-3.2, ИПК-3.3, ИПК-3.4, ИПК-4.4. Ответы на вопросы первой части даются путем выбора из списка предложенных.

Примерный перечень теоретических вопросов

1. Опишите типологизацию направления «Обработка естественного языка»
2. Опишите методы векторизации слов? Какие методы существуют их преимущества и недостатки
3. Опишите принцип работы метода векторизации word2vec, опишите его преимущества и недостатки
4. Опишите принцип работы метода векторизации FastText, опишите его преимущества и недостатки
5. Дайте определение расстояния. В каких методах обработки естественного языка применяется данный подход?
6. Автоматическая обработка естественного языка, ее цели и задачи. Предмет и объект данной области знаний.
7. Принцип работы контекстно свободных грамматик
8. Опишите принцип работы Word Embeddings: независимые от контекста представления слов
9. Опишите принцип работы Word Embeddings: зависимые от контекста представления слов
10. Обучение интеллектуальных систем. Виды обучающихся интеллектуальных систем
11. Предобработка данных, типы предобработки, ее особенности.
12. Задачи и методы генерации текстов. Файнтьюн обученной модели.
13. Принципы работы нейронной сети GPT-3.
14. Опишите методы извлечения фактов из текста. Опишите их преимущества и недостатки.

15. Дайте определение корпусной лингвистики?

16. Что означает формула IPM и в каких случаях она применяется?

17. Приведите типологизацию НКРЯ

18. Дайте определение контекстно-свободных грамматик?

19. Опишите формальные метрики точности работы классификаторов. В чем преимущества и недостатки формальных метрик?

Вторая часть содержит один вопрос, проверяющий ИПК-3.2, ИПК-3.3, ИПК-3.4, ИПК-4.4. Ответы на вопросы второй части предполагают решение задач и краткую интерпретацию полученных результатов.

Примерный перечень практических вопросов:

1. С помощью диалектного корпуса найти территории, где употребляются слова:

а) евоный

б) худо

в) пошто

2. Когда впервые было употреблено слово «диверсификация», в какой сфере функционирования данное слово употребляется чаще всего.

4. Основываясь на формуле TF-IDF найдите ключевые слова в новостных текстах рубрики «Спорт»

Третья часть предусматривает защиту индивидуального или группового проекта по одной из областей NLP.

Результаты экзамена определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Критерии оценки экзамена обусловлены логической демонстрацией приобретенных компетенций в соответствии с текущей программой. Демонстрация предусматривает уверенное использование терминологии, понимание и корректное использование математического аппарата, предусматривает корректность написания кода, его понимание и корректное использование в нем математических методов. Отметка «хорошо» выставляется за счет демонстрации полученных компетенций, владение и понимание кода, теоретических аспектов его применения в практике работы с текстовыми массивами данных допускаются недочеты в понятийном аппарате математики. Отметка «удовлетворительно» позволяет допустить ошибки в разработке кода, но учитывет последовательную логику изложения структуры кода, его интерпретацию, связь теоретических аспектов лингвистики и математики, демонстрация понимания хода обработки текста. Минимальный порог оценки «отлично» составляет 90-100 баллов, хорошо 75-89, удовлетворительно «55-74» ниже 55 – «неудовлетворительно»

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» - https://lms.tsu.ru/course/view.php?id=12998

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

в) План семинарских / практических занятий по дисциплине.

Семинар №1

1. Парсинг и структуризация данных

2. Разработка корпуса

Семинар №2

1. Предобработка текстовых данных

2. Лемматизация текстов при помощи mystem/pymorpy/udipipe

Семинар №3

1. Библиотека NLTK для векторизации и анализа текстовых массивов данных

2. Векторизация текстов. Принципы, методы

3. Составление словарей, n-грамы

Семинар №4

1. Transforers: генерация текста

5. Transforers: вопросно-ответные системы

Семинар №6

1. Сокращение пространства признаков

2. Визуализация и анализ текстовых данных

Подготовка к проведению лабораторных работ начинается в начале теоретического изложения изучаемой темы и продолжается по ходу её изучения при освоении материала на занятиях в рамках практических заданий и работе над ним в ходе самостоятельной подготовки дома и в библиотеках. Для качественного выполнения лабораторных работ студентам необходимо:

1) повторить теоретический материал по конспекту и учебникам;

2) ознакомиться с описанием лабораторной работы;

3) в специальной тетради для лабораторных работ записать название и номер работы, перечень необходимого программного обеспечения, подготовить алгоритм или код;

4) выяснить цель работы, четко представить себе поставленную задачу и способы её достижения, продумать ожидаемые результаты опытов;

5) ответить устно или письменно на контрольные вопросы по изучаемой теме или решить ряд задач;

6) изучить порядок выполнения лабораторной работы. Подготовить среду выполнения кода к работе. После проверки правильности алгоритма работы программы преподавателем можно начинать выполнение лабораторной работы.

д) Методические указания по организации самостоятельной работы студентов.

Формы самостоятельной работы студентов разнообразны. Они включают в себя:

– изучение и систематизацию практических и теоретических примеров в рамках выполнения текущих заданий по предмету;

– изучение учебной, научной и методической литературы, материалов периодических изданий с привлечением электронных средств официальной, статистической, периодической и научной информации;

– подготовку докладов и презентаций, написание программного кода и его отладка;

– участие в работе студенческих конференций, комплексных научных исследованиях.

Самостоятельная работа приобщает студентов к научному творчеству, поиску и решению актуальных современных проблем.

Примеры для самостоятельной работы студентов:

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Jurafsky J., James H M. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Model. , 2024. Вып. 3. 591 с.

– Васильев Ю. Обработка естественного языка. Python и spaCy на практике. : Питер, 2024. 358 с.

– Хобсон Л., Ханнес Х., Коул Х. Обработка естественного языка в действии. : Питер, 2024. 597 с.

– Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др.–М.: МИЭМ, 2011.
– Когнитивная и компьютерная лингвистика / Ред.: Р.Г.Бухараев, В.Д.Соловьев, Д.Ш.Сулейманов. - Казань: КГУ, 1994. - 112 с
– Баранов А. Н. Введение в прикладную лингвистику. М., 2001. URL: https://dislyget.ru/index.php?r=item/view&id=21065

б) дополнительная литература:

– 1. Гольдберг Й. Нейросетевые методы в обработке естественного языка. : Litres, 2022. 284 с.
– Пойнтер Я. Программируем с PyTorch: Создание приложений глубокого обучения. : Питер, 2024. 289 с.
1. Thakur V., Tickoo A. Text2Gender: A Deep Learning Architecture for Analysis of Blogger's Age and Gender // 2023.
– Thomas A. Natural Language Processing with Spark NLP: Learning to Understand Text at Scale. : O'Reilly Media, Inc., 2020. 367 с.
– Vajjala S. и др. Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems. : O'Reilly Media, Inc., 2020. 455 с.
– Tunstall L., Werra L. von, Wolf T. Natural Language Processing with Transformers. : O'Reilly Media, Inc., 2022. 409 с.
– Rothman D. Transformers for Natural Language Processing and Computer Vision: Explore Generative AI and Large Language Models with Hugging Face, ChatGPT, GPT-4V, and DALL-E 3. : Packt Publishing Ltd, 2024. 731 с.
– Rothman D. Transformers for Natural Language Processing: Build, train, and fine-tune deep neural network architectures for NLP with Python, Hugging Face, and OpenAI's GPT-3, ChatGPT, and GPT-4. : Packt Publishing Ltd, 2022. 603 с.
– Rothman D. Transformers for Natural Language Processing: Build Innovative Deep Neural Network Architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and More. : Packt Publishing, Limited, 2021. 384 с.
– Rao D., McMahan B. Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning. : O'Reilly Media, Inc., 2019. 256 с.

в) ресурсы сети Интернет:

– открытые онлайн-курсы;
– Python документация https://docs.python.org/3/index.html;
– Репозиторий HuggingFace https://huggingface.co/, GitHub https://github.com/;
– Google Colab https://colab.google/ и/или Kagle https://www.kaggle.com/;
– публично доступные облачные технологии (Google Docs, Яндекс диск и т.п.);
– язык программирования Python https://www.python.org/.

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

– Текстовые редакторы; MS Ofaice или любой другой, notepad++ или Sublime Text
– публично доступные облачные технологии (Google Docs, Яндекс диск и т.п.);
– язык программирования и Python ver. > 3 и jupyter notebook (anaconda).
- СУБД PostgreSQL
- ПО Mystem
- Создание окружения: gitBash, Docker,

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

в) профессиональные базы данных *(при наличии)*:
– Национальный корпус русского языка - https://ruscorpora.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.

Лаборатории, оборудованные компьютерами (не ниже i5, RAM 16Gb), проектором

Аудитории для проведения занятий лекционного и семинарского типа индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации в смешанном формате («Актру»).

15. Информация о разработчиках

Степаненко Андрей Александрович, НИ Томский государственный университет, старший преподаватель кафедры общей, компьютерной и когнитивной лингвистики.