

**Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»**

УТВЕРЖДЕНО
Заместитель директора

Ю.О. Соболев

	Рабочая программа дисциплины (модуля)
по дисциплине:	Задачи генерации в NLP
по направлению:	Прикладная математика и информатика
профиль подготовки:	Науки о данных
	центр дополнительного, дополнительного профессионального и онлайн-образования "Пуск"
	центр дополнительного, дополнительного профессионального и онлайн-образования "Пуск"
курс:	2
квалификация:	магистр

Семестр, формы промежуточной аттестации: 4 (весенний) - Дифференцированный зачет

Аудиторных часов: 22 всего, в том числе:

лекции: 2 час.

семинары: 20 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 113 час.

Всего часов: 135, всего зач. ед.: 3

Программу составили:

Н.В. Сидоров, преподаватель

К.А. Лапин, старший методист

Программа обсуждена на заседании центра дополнительного, дополнительного профессионального и
онлайн-образования "Пуск" 01.03.2025

1. Цели и задачи

Цель дисциплины

- познакомить обучающихся с компьютерной лингвистикой (КЛ), как научно-практическим направлением, его краткой историей, познакомить с задачами, связанными с обработкой основных лингвистических и текстовых параметров языковых произведений, а также с методами и технологиями, используемыми в рамках компьютерной лингвистики.

Задачи дисциплины

- создать представление о компьютерной лингвистике как новейшей научно-практической области исследований, ее возникновении в контексте смежных наук и ее современной организации;
- познакомить обучающихся с основными лингвистическими технологиями, реализующими анализ предложения (текста) по уровням лингвистической разметки и основными приемами автоматической генерации текстов;
- познакомить обучающихся с основными типами ресурсов, создающимися и используемыми компьютерными программами для решения конкретных задач в исследовательских целях, при разработке лингвистических технологий и в приложениях;
- соединить интуитивные и традиционные представления о свойствах естественно-языковых текстов со способами их формализации и моделирования в работах по компьютерной лингвистике;
- выработать у обучающихся элементарные практические навыки по применению компьютерно-лингвистических методов к языковому материалу и использованию лингвистических технологий.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий	УК-1.1 Анализирует проблемную ситуацию как систему, выявляя ее составляющие и связи между ними
	УК-1.2 Осуществляет поиск вариантов решения поставленной проблемной ситуации на основе доступных источников информации
	УК-1.3 Разрабатывает стратегию достижения поставленной цели как последовательность шагов, предвидя результат каждого из них и оценивая их влияние на внешнее окружение планируемой деятельности и на взаимоотношения участников этой деятельности
УК-5 Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия	УК-5.1 Способен выявлять специфику философских и научных традиций основных мировых культур
	УК-5.2 Способен определять теоретическое и практическое значение культурно-языкового фактора при взаимодействии различных философских и научных традиций
ОПК-2 Способен совершенствовать и реализовывать новые математические	ОПК-2.1 Имеет представление о современном состоянии математических исследований в рамках тематической области своей профессиональной деятельности
	ОПК-2.2 Способен оценить актуальность и практическую значимость прикладных математических исследований в своей профессиональной области

методы решения прикладных задач

ОПК-2.3 Владеет профессиональной терминологией, используемой в современной научно-технической литературе, обладает навыками устного и письменного изложения результатов научной деятельности в рамках профессиональной коммуникации

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- специфические особенности задач генерации текстов;
- формулировки постановок задач генерации текстов и метрики оценки качества;
- основные модели для решения задач генерации.

уметь:

- решать задачи средствами huggingface transformers.

владеть:

- state of the art инструментами NLP для решения задач генерации текстов.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Особенности использования трансформеров для генерации текстов	2	8		23
2	Задача перевода текста		4		28
3	Задача генерации текстов		4		29
4	Задача абстрактивной саммаризации		2		15
5	Использование LLM для генерации текстов		2		18
Итого часов		2	20		113
Подготовка к экзамену		0 час.			
Общая трудоёмкость		135 час., 3 зач.ед.			

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 4 (Весенний)

1. Особенности использования трансформеров для генерации текстов

Виды предобученных языковых моделей на основе архитектуры Transformer. Отличие задачи генерации от задачи извлечения информации. Задача text retrieval. Методы поиска информации. Задача языкового моделирования.

2. Задача перевода текста

Задача перевода с одного языка на другой. Метрики оценки перевода. Задача перевода на малоресурсные языки. Нейросетевые модели для перевода. Стратегии генерации текста.

3. Задача генерации текстов

Задача text-style-transfer. Контекст языковых моделей и его изменение. Основы промпт-инжиниринга. Бенчмарки для сравнения языковых моделей.

4. Задача абстрактивной саммаризации

Задача суммаризации. Метрики для оценки суммаризированного текста. Нейросетевые модели для суммаризации текстов. RLHF для задачи суммаризации.

5. Использование LLM для генерации текстов

Обзор инструктивных моделей. PEFT-методы для обучения моделей. Методы борьбы с галлюцинациями в генеративных моделях. Прототипирование решений с использованием LLM. Альтернативные архитектуры LLM.

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Обучение проходит с использованием дистанционных образовательных технологий.

6. Перечень рекомендуемой литературы

Основная литература

1. Машинное обучение: новый искусственный интеллект [Текст]/Э. Алпайдин, -М., Изд. группа "Точка", 2017

Дополнительная литература

1. Python и машинное обучение [Текст], крайне необходимое издание по новейшей предсказательной аналитике для более глубокого понимания методологии машинного обучения/С. Рашка, -М., ДМК Пресс, 2017

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

1. Введение в анализ данных с помощью Pandas - <https://habr.com/ru/post/196980/>
2. Learn Git Branching - https://learngitbranching.js.org/?locale=ru_RU
3. <https://cs231n.github.io/python-numpy-tutorial/>
4. <https://www.deeplearningbook.org/>
5. <https://uproger.com/10-bibliotek-python-dlya-mashinnogo-obucheniya/?ref=vc.ru>

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

1. Google Drive для доступа к материалам курса
2. Zoom
3. Ноутбук для участия в интерактивных занятиях

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Студент, изучающий дисциплину, должен с одной стороны, овладеть общим понятийным аппаратом, а с другой стороны, должен научиться применять теоретические знания на практике. В результате изучения дисциплины студент должен знать основные определения дисциплины, уметь применять полученные знания для решения различных задач.

Успешное освоение курса требует:

- посещения всех занятий, предусмотренных учебным планом по дисциплине;
- ведения конспекта занятий;
- напряжённой самостоятельной работы студента.

Самостоятельная работа включает в себя:

- чтение рекомендованной литературы;
- проработку учебного материала, подготовку ответов на вопросы, предназначенных для самостоятельного изучения;
- решение задач, предлагаемых студентам на занятиях;
- подготовку к выполнению заданий текущей и промежуточной аттестации.

Показателем владения материалом служит умение без конспекта отвечать на вопросы по темам дисциплины.

Важно добиться понимания изучаемого материала, а не механического его запоминания. При затруднении изучения отдельных тем, вопросов, следует обращаться за консультациями к преподавателю.

Возможен промежуточный контроль знаний студентов в виде решения задач в соответствии с тематикой занятий.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению:	Прикладная математика и информатика
профиль подготовки:	Науки о данных центр дополнительного, дополнительного профессионального и онлайн-образования "Пуск" центр дополнительного, дополнительного профессионального и онлайн-образования "Пуск"
курс:	<u>2</u>
квалификация:	магистр

Семестр, формы промежуточной аттестации: 4 (весенний) - Дифференцированный зачет

Разработчики:

Н.В. Сидоров, преподаватель
К.А. Лапин, старший методист

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий	УК-1.1 Анализирует проблемную ситуацию как систему, выявляя ее составляющие и связи между ними
	УК-1.2 Осуществляет поиск вариантов решения поставленной проблемной ситуации на основе доступных источников информации
	УК-1.3 Разрабатывает стратегию достижения поставленной цели как последовательность шагов, предвидя результат каждого из них и оценивая их влияние на внешнее окружение планируемой деятельности и на взаимоотношения участников этой деятельности
УК-5 Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия	УК-5.1 Способен выявлять специфику философских и научных традиций основных мировых культур
	УК-5.2 Способен определять теоретическое и практическое значение культурно-языкового фактора при взаимодействии различных философских и научных традиций
ОПК-2 Способен совершенствовать и реализовывать новые математические методы решения прикладных задач	ОПК-2.1 Имеет представление о современном состоянии математических исследований в рамках тематической области своей профессиональной деятельности
	ОПК-2.2 Способен оценить актуальность и практическую значимость прикладных математических исследований в своей профессиональной области
	ОПК-2.3 Владеет профессиональной терминологией, используемой в современной научно-технической литературе, обладает навыками устного и письменного изложения результатов научной деятельности в рамках профессиональной коммуникации

2. Показатели оценивания компетенций

В результате изучения дисциплины «Задачи генерации в NLP» обучающийся должен:

знать:

- специфические особенности задач генерации текстов;
- формулировки постановок задач генерации текстов и метрики оценки качества;
- основные модели для решения задач генерации.

уметь:

- решать задачи средствами huggingface transformers.

владеть:

- state of the art инструментами NLP для решения задач генерации текстов.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

Примерный перечень вопросов для текущего контроля:

1. Почему модель BERT обученная с помощью masked language modeling не подходит для задач генерации текстов
2. Классификация текстов. Классические (BOW, TF-IDF) и нейросетевые подходы (на основе свёрточных, рекуррентных сетей,
3. AlchemyAPI
4. Expert System S.p.A.
5. General Architecture for Text Engineering (GATE)

6. Modular Audio Recognition Framework
7. MontyLingua
8. Natural Language Toolkit (NLTK)

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

Примерные вопросы для дифференцированного зачета:

1. Архитектура трансформер. Особенности использования трансформеров для задач генерации текстов.
2. Задача машинного перевода. Метрики оценки качества.
3. Распознавание речи
4. Обработка текста
5. Извлечение информации.
7. Анализ информации.
8. Генерация текста и речи.
9. Автоматический пересказ
10. Машинный перевод

Критерии оценивания

- оценка «отлично (10)» выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений
- оценка «отлично (9)» выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений
- оценка «отлично (8)» выставляется студенту, показавшему всесторонние систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, и правильное обоснование принятых решений
- оценка «хорошо (7)» выставляется студенту, если он твердо знает материал, грамотно и, по существу, излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «хорошо (6)» выставляется студенту, если он знает материал, грамотно и, по существу, излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «хорошо (5)» выставляется студенту, если он знает материал, и, по существу, излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «удовлетворительно (4)» выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;
- оценка «удовлетворительно (3)» выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет фрагментарно основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;
- оценка «неудовлетворительно (2)» выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач
- оценка «неудовлетворительно (1)» выставляется студенту, который не знает формулировок основных понятий дисциплины.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Промежуточная аттестация проходит в формате дифференцированного зачета на lms.

Время отведенное на дифференцированный зачет: 2 академических часа.