

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

УТВЕРЖДЕНО
Заведующий кафедрой

В.М. Говорун

	Рабочая программа дисциплины (модуля)
по дисциплине:	Основы алгоритмов обработки естественных языков NLP
по направлению:	Прикладные математика и физика
профиль подготовки:	Системная и синтетическая биология Физтех-школа Биологической и Медицинской Физики кафедра системной и синтетической биологии
курс:	4
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 8 (весенний) - Дифференцированный зачет

Аудиторных часов: 30 всего, в том числе:

лекции: 0 час.

семинары: 30 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 15 час.

Всего часов: 45, всего зач. ед.: 1

Программу составил: В.М. Говорун, д-р биол. наук, профессор

Программа обсуждена на заседании кафедры системной и синтетической биологии 07.03.2025

Аннотация

Курс разработан для бакалавров, не имеющих глубоких знаний в машинном обучении и программировании, поэтому материал представлен в доступной форме. В рамках курса рассматриваются основные принципы машинного обучения, ключевые методы NLP и их использование для решения задач в медицинской сфере. Обучение включает как теоретические занятия, так и практические лабораторные работы, где студенты смогут применять изученные методы на реальных примерах из персонализированной медицины.

В завершение курса студенты выполняют курсовой проект, демонстрирующий применение NLP для анализа медицинских данных.

1. Цели и задачи

Цель дисциплины

- ознакомление студентов с основами алгоритмов обработки естественного языка (NLP) и их применением в медицине.

Задачи дисциплины

1. Изучить основные концепции машинного обучения и NLP, познакомиться с ключевыми терминами и методами, включая классификацию текстов, семантический анализ и извлечение информации.
2. Освоить базовые принципы работы с языком программирования Python в контексте анализа данных, включая предобработку текстов и применение специализированных библиотек.
3. Выполнить практические задания по медицинскому NLP, связанные с анализом медицинских текстов, обработкой записей и решением других прикладных задач в здравоохранении.
4. Изучить методы оценки качества моделей NLP и их эффективности при решении задач медицинского анализа текстов.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-1 Способен применять фундаментальные знания, полученные в области физико-математических и (или) естественных наук, и использовать их в профессиональной деятельности	ОПК-1.1 Способен анализировать поставленную задачу, намечать пути ее решения
ОПК-4 Способен осуществлять сбор и обработку научно-технической и (или) технологической информации для решения фундаментальных и прикладных задач	ОПК-4.1 Владеет методами научного поиска и интеллектуального анализа информации при решении задач профессиональной деятельности
	ОПК-4.2 Знает основные источники научно-технической и (или) технологической информации в области профессиональной деятельности
	ОПК-4.4 Владеет навыками работы с компьютером и компьютерными сетями с целью получения, хранения и обработки научной (технической, технологической) информации
ПК-2 Способен анализировать полученные в ходе научно-исследовательской работы данные и делать научные выводы (заключения)	ПК-2.2 Умеет находить ключевые параметры, определяющие изучаемое явление, и производить численные оценки по порядку величины
	ПК-2.3 Способен представлять научные утверждения, их обоснования и доказательства, научные проблемы и их решения ясно и точно в терминах, понятных для профессиональной аудитории, в письменной и устной форме
ПК-4 Способен критически оценивать применимость используемых методик и	ПК-4.2 Знает источники происхождения и умеет производить оценку погрешности измерений и достоверности экспериментальных результатов

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основные методы машинного обучения, включая классификацию, регрессию и кластеризацию;
- ключевые концепции обработки естественного языка, анализа структуры текста;
- особенности медицинских текстовых данных и их отличия от других видов текстовой информации;
- основные принципы работы современных NLP-моделей, включая трансформеры (BERT, GPT) и их применение в медицине;
- основные требования к защите и обработке медицинских данных.

уметь:

- использовать Python для обработки медицинских текстов и применять библиотеки для анализа данных;
- разрабатывать и настраивать простые NLP-модели для решения задач классификации, извлечения информации и анализа медицинских текстов;
- оценивать качество работы моделей с использованием метрик точности, полноты и других показателей;
- анализировать медицинские тексты и извлекать из них полезную информацию с помощью методов NLP.

владеть:

- навыками работы с текстовыми данными в Python, включая их предобработку, анализ и визуализацию;
- навыками применения алгоритмов машинного обучения для работы с медицинскими текстами;
- навыками создания и тестирования простейших больших языковых моделей, адаптированных для анализа медицинских 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	Введение в машинное обучение и NLP		5		2
2	Python для анализа медицинских текстов		5		3
3	Большие языковые модели и их применение в медицине		5		3
4	Извлечение информации и анализ медицинских текстов		5		3
5	Оценка качества моделей и их внедрение		5		2
6	Этика, правовые аспекты и будущее медицинского NLP		5		2
Итого часов			30		15
Подготовка к экзамену		0 час.			
Общая трудоёмкость		45 час., 1 зач.ед.			

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

Семестр: 8 (Весенний)

1. Введение в машинное обучение и NLP

Основные методы машинного обучения: классификация, регрессия, кластеризация. Ключевые задачи NLP и их применение в медицине.

Машинное обучение: основные задачи и методы. Классификация, регрессия, кластеризация в анализе данных. Основные концепции NLP: обработка текстов, извлечение информации, генерация текста. Применение NLP в медицине: анализ медицинских записей, диагностика заболеваний, поддержка врачебных решений.

2. Python для анализа медицинских текстов

Основы работы с данными в Python. Основные библиотеки для NLP: Pandas, NLTK, SpaCy. Предобработка текстов: токенизация, лемматизация, POS-теггинг. Основы Python: работа с текстовыми и табличными данными. Библиотеки для анализа текстов: Pandas для обработки данных, NLTK и SpaCy для NLP. Предобработка медицинских текстов: токенизация, стемминг, лемматизация. POS-теггинг и синтаксический разбор в анализе клинических данных.

3. Большие языковые модели и их применение в медицине

Развитие NLP: от Bag-of-Words до трансформеров. Архитектура и особенности моделей BERT, GPT и их медицинских вариаций. Аннотация и обработка медицинских данных. История NLP: от классических методов (Bag-of-Words, TF-IDF) до трансформеров. Архитектура BERT, GPT: принципы работы, особенности обучения. Медицинские вариации моделей: BioBERT, ClinicalBERT. Аннотация и обработка медицинских данных: подготовка датасетов, маркировка, валидация.

4. Извлечение информации и анализ медицинских текстов

Методы извлечения ключевой информации. Автоматическая классификация и анализ медицинских записей. Использование NLP в биоинформатике и медицинских исследованиях. Извлечение информации: распознавание именованных сущностей, семантический анализ. Автоматическая классификация медицинских документов. Анализ электронных медицинских записей: выявление симптомов, прогнозирование заболеваний. NLP в биоинформатике: анализ генетических данных, обработка научных публикаций.

5. Оценка качества моделей и их внедрение

Метрики оценки NLP-моделей. Улучшение и адаптация моделей к медицинским данным. Интеграция NLP-решений в клинические и исследовательские процессы. Метрики качества: точность, полнота, F1-мера, AUC-ROC. Методы оптимизации моделей: дообучение, настройка гиперпараметров. Внедрение NLP-моделей: автоматизация медицинской документации, анализ клинических записей. Интеграция в медицинские информационные системы.

6. Этика, правовые аспекты и будущее медицинского NLP

Конфиденциальность и защита медицинских данных. Этика работы с медицинскими текстами. Перспективы развития NLP и AI в медицине. Правовые аспекты: защита персональных данных, HIPAA, GDPR. Этика в медицинском NLP: риски автоматизированного анализа, врачебные ошибки. Перспективы развития: мультимодальные модели, персонализированная медицина, автоматизированные медицинские консультации.

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

Учебные аудитории, оснащенные мультимедийным оборудованием (экран, проектор, аудио и видеоаппаратура, ноутбук с подключением к сети «Интернет», микрофоны).
Персональные Компьютеры (Ноутбуки) студентов для выполнения практических заданий и выполнения домашней работы.

6.Перечень рекомендуемой литературы

Основная литература

Основная литература предоставлена базовой кафедрой

1. Панов А. И. Введение в методы машинного обучения с подкреплением: учебное пособие / А. И. Панов. — Москва: МФТИ, 2019. — 52 с.
2. De Marchi L., Mitchell L. Hands-On Neural Networks: Learn How to Build and Train Your First Neural Network Model Using Python / L. De Marchi, L. Mitchell. — Birmingham: Packt Publishing, 2019. — ISBN 978-1-78899-259-6, ISBN 978-1-78899-988-5.
3. Рашка С. Python и машинное обучение. – Litres, 2022.

Дополнительная литература

Литература предоставлена базовой кафедрой

1. Литвин А. А. и др. Новые возможности искусственного интеллекта в медицине: описательный обзор //Проблемы здоровья и экологии. – 2024. – Т. 21. – №. 1. – С. 7-17.
2. Керимов К. Ф., Мухсинов Ш. Ш., Вохдатхужаев А. В. ИЗВЛЕЧЕНИЕ КЛИНИЧЕСКИ ЗНАЧИМЫХ ФЕНОТИПОВ ИЗ ЗАПИСЕЙ ЭЛЕКТРОННЫХ МЕДИЦИНСКИХ КАРТ //Journal of new century innovations. – 2023. – Т. 27. – №. 5. – С. 212-215.
3. Боброва Е. В. и др. Генерация врачебных заключений и классификация по Bethesda с использованием глубокого обучения //International Journal of Open Information Technologies. – 2023. – Т. 11. – №. 10. – С. 119-129.
4. Balbek K., Melerzanov A. Modern approaches of mapping electronic health records (ehr) to human phenotype ontology (hpo) using advanced language models. Health care Standardization Problems, 2023.
5. Камолова Д. П. ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В ЮРИСПРУДЕНЦИИ: ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ //Состав редакционной коллегии и организационного комитета. – 2023.

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

Hugging Face. Доступ по ссылке: <https://huggingface.co>. Дата обращения: [10.04.2024].

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

Для части занятий потребуется Яндекс.Телемост. Потребуется Яндекс.Диск для доступа к материалам курса. Потребуется наличие ноутбуков у студентов для участия в интерактивных упражнениях

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.

ПРИЛОЖЕНИЕ

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Прикладные математика и физика
профиль подготовки: Системная и синтетическая биология
Физтех-школа Биологической и Медицинской Физики
кафедра системной и синтетической биологии
курс: 4
квалификация: бакалавр

Семестр, формы промежуточной аттестации: 8 (весенний) - Дифференцированный зачет

Разработчик: В.М. Говорун, д-р биол. наук, профессор

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-1 Способен применять фундаментальные знания, полученные в области физико-математических и (или) естественных наук, и использовать их в профессиональной деятельности	ОПК-1.1 Способен анализировать поставленную задачу, намечать пути ее решения
ОПК-4 Способен осуществлять сбор и обработку научно-технической и (или) технологической информации для решения фундаментальных и прикладных задач	ОПК-4.1 Владеет методами научного поиска и интеллектуального анализа информации при решении задач профессиональной деятельности
	ОПК-4.2 Знает основные источники научно-технической и (или) технологической информации в области профессиональной деятельности
	ОПК-4.4 Владеет навыками работы с компьютером и компьютерными сетями с целью получения, хранения и обработки научной (технической, технологической) информации
ПК-2 Способен анализировать полученные в ходе научно-исследовательской работы данные и делать научные выводы (заключения)	ПК-2.2 Умеет находить ключевые параметры, определяющие изучаемое явление, и производить численные оценки по порядку величины
	ПК-2.3 Способен представлять научные утверждения, их обоснования и доказательства, научные проблемы и их решения ясно и точно в терминах, понятных для профессиональной аудитории, в письменной и устной форме
ПК-4 Способен критически оценивать применимость используемых методик и методов	ПК-4.2 Знает источники происхождения и умеет производить оценку погрешности измерений и достоверности экспериментальных результатов
	ПК-4.3 Способен обосновать причинно-следственные отношения используемых понятий и моделей

2. Показатели оценивания компетенций

В результате изучения дисциплины «Основы алгоритмов обработки естественных языков NLP» обучающийся должен:

знать:

- основные методы машинного обучения, включая классификацию, регрессию и кластеризацию;
- ключевые концепции обработки естественного языка, анализа структуры текста;
- особенности медицинских текстовых данных и их отличия от других видов текстовой информации;
- основные принципы работы современных NLP-моделей, включая трансформеры (BERT, GPT) и их применение в медицине;
- основные требования к защите и обработке медицинских данных.

уметь:

- использовать Python для обработки медицинских текстов и применять библиотеки для анализа данных;
- разрабатывать и настраивать простые NLP-модели для решения задач классификации, извлечения информации и анализа медицинских текстов;
- оценивать качество работы моделей с использованием метрик точности, полноты и других показателей;
- анализировать медицинские тексты и извлекать из них полезную информацию с помощью методов NLP.

владеть:

- навыками работы с текстовыми данными в Python, включая их предобработку, анализ и визуализацию;
- навыками применения алгоритмов машинного обучения для работы с медицинскими текстами;
- навыками создания и тестирования простейших больших языковых моделей, адаптированных для анализа медицинских текстов.

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

1. Основные методы машинного обучения: классификация, регрессия, кластеризация.
2. Ключевые задачи NLP и их применение в медицине.
3. Машинное обучение: основные задачи и методы.
4. Классификация, регрессия, кластеризация в анализе данных.
5. Основные концепции NLP: обработка текстов, извлечение информации, генерация текста.
6. Применение NLP в медицине: анализ медицинских записей, диагностика заболеваний, поддержка врачебных решений.
7. Основы работы с данными в Python.
8. Основные библиотеки для NLP: Pandas, NLTK, SpaCy.
9. Предобработка текстов: токенизация, лемматизация, POS-теггинг.
10. Основы Python: работа с текстовыми и табличными данными.

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

Вопросы к дифференцированному зачету:

1. Библиотеки для анализа текстов: Pandas для обработки данных, NLTK и SpaCy для NLP.
2. Предобработка медицинских текстов: токенизация, стемминг, лемматизация.
3. POS-теггинг и синтаксический разбор в анализе клинических данных.
4. Извлечение информации и анализ медицинских текстов
5. Архитектура и особенности моделей BERT, GPT и их медицинских вариаций.
6. Аннотация и обработка медицинских данных.
7. Архитектура BERT, GPT: принципы работы, особенности обучения.
8. Медицинские вариации моделей: BioBERT, ClinicalBERT.
9. Аннотация и обработка медицинских данных: подготовка датасетов, маркировка, валидация.
10. Развитие NLP: от Bag-of-Words до трансформеров.

Критерии оценивания

Оценка отлично 10 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины, проявляющему интерес к данной предметной области, продемонстрировавшему умение уверенно и творчески применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

Оценка отлично 9 баллов - 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 8 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, правильное обоснование принятых решений, с некоторыми недочетами.

Оценка хорошо 7 баллов - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но недостаточно грамотно обосновывает полученные результаты.

Оценка хорошо 6 баллов - 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 5 баллов - выставляется студенту, если он в основном знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач достаточно большое количество неточностей.

Оценка удовлетворительно 4 балла - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он освоил основные разделы учебной программы, необходимые для дальнейшего обучения, и может применять полученные знания по образцу в стандартной ситуации.

Оценка удовлетворительно 3 балла - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, допускающему ошибки в формулировках базовых понятий, нарушения логической последовательности в изложении программного материала, слабо владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и с трудом применяет полученные знания даже в стандартной ситуации.

Оценка неудовлетворительно 2 балла - выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных принципов и не умеет использовать полученные знания при решении типовых задач.

Оценка неудовлетворительно 1 балл - выставляется студенту, который не знает основного содержания учебной программы дисциплины, допускает грубейшие ошибки в формулировках базовых понятий дисциплины и вообще не имеет навыков решения типовых практических задач.

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

Аттестация по дисциплине осуществляется в форме дифференцированного зачета. При проведении устного дифференцированного зачета обучающемуся предоставляется 45 минут на подготовку. Опрос обучающегося по билету при устном ответе не должен превышать одного астрономического часа.