

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

УТВЕРЖДЕНО
исполнительный директор

М.А. Смирнова

	Рабочая программа дисциплины (модуля)
по дисциплине:	Основы алгоритмов обработки естественного языка в индустрии здравоохранения
по направлению:	Биотехнология
профиль подготовки:	Системная и синтетическая биология Физтех-школа Биологической и Медицинской Физики кафедра системной и синтетической биологии
курс:	1
квалификация:	магистр

Семестры, формы промежуточной аттестации:

1 (осенний) - Зачет
2 (весенний) - Экзамен

Аудиторных часов: 60 всего, в том числе:

лекции: 0 час.
семинары: 60 час.
лабораторные занятия: 0 час.

Самостоятельная работа: 135 час.

Подготовка к экзамену: 30 час.

Всего часов: 225, всего зач. ед.: 5

Программу составил: К.Д. Балбек

Программа обсуждена на заседании кафедры системной и синтетической биологии 11.04.2024

Аннотация

Целью данной дисциплины является получение студентами представления об алгоритмах обработки естественного языка (NLP) в медицине. Курс также рассчитан для студентов, не имеющих глубоких знаний в машинном обучении или программировании на Python. Курс включает в себя основы машинного обучения, основные концепции и техники NLP, а также их приложение для решения задач в медицинской сфере. Курс сочетает теоретические занятия с практическими лабораторными работами с целью применения студентами изученных методов для решения аналитических задач в персонализированной биомедицине. В конце годовой программы студентам будет предложено защитить курсовой проект, являющийся примером применения NLP при решении биомедицинских задач.

1. Цели и задачи

Цель дисциплины

- обеспечение понимания основ машинного обучения и NLP;
- ознакомление с базовыми принципами и методами машинного обучения и обработки естественного языка, а также их применением в медицинской индустрии;
- развитие базовых навыков программирования на Python, необходимых для работы с NLP-библиотеками и инструментами анализа данных;
- приобретение практического опыта в медицинском NLP на реальных задачах по обработке медицинской информации.

Задачи дисциплины

- изучить ключевые концепции машинного обучения и NLP, освоить основные термины и методы, такие как классификация текстов, семантический анализ и извлечение информации;
- изучить основные подходы по использованию языка программирования Python для обработки данных - написание кода для предобработки данных, использование библиотек, и реализация базовых моделей машинного обучения;
- реализовать проекты по медицинскому NLP, направленные на анализ медицинских записей, а также на другие задачи, связанные с обработкой текстов в здравоохранении;
- научиться оценивать эффективность и точность моделей NLP в контексте медицинских приложений.

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
ОПК-2 Имеет представление об актуальных проблемах науки и техники в области своей профессиональной деятельности, способен	ОПК-2.1 Имеет представление о современном состоянии исследований в рамках тематической области своей профессиональной деятельности
	ОПК-2.2 Способен оценивать актуальность исследований в области своей профессиональной деятельности и их практическую значимость

на научном языке формулировать профессиональные задачи	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
ОПК-3 Способен выбирать и (или) разрабатывать подходы к решению типовых и новых задач в области профессиональной деятельности, учитывая особенности и ограничения различных методов решения	ОПК-3.1 Способен анализировать задачу, планировать пути решения, предлагать и комбинировать способы решения
	ОПК-3.2 Способен использовать исследовательские методы при решении новых задач, применяя знания в различных областях науки (техники)
	ОПК-3.3 Владеет аналитическими и вычислительными методами решения, понимает и учитывает на практике границы применимости получаемых решений
	ОПК-3.4 Способен к профессиональной эксплуатации и модернизации современного технологического оборудования для осуществления биотехнологических процессов
	ОПК-3.5 Владеет навыками проектирования новых биотехнологических решений для поставленных научно-технических и технологических задач
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.2 Способен использовать специализированные знания фундаментальных разделов математики, физики, химии и биологии для постановки и решения научно-исследовательских задач в области биоинженерии и биоинформатики
	ПК-1.3 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.4 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
	ПК-1.5 Способен создавать программные средства и базы данных, используемые в биоинженерии и биоинформатике
ПК-3 Способен профессионально работать с исследовательским и испытательным оборудованием (приборами и установками, специализированными пакетами прикладных программ) в избранной предметной области	ПК-3.1 Понимает принципы работы используемого оборудования (специализированных пакетов прикладных программ)
	ПК-3.2 Способен проводить эксперимент (моделирование) с использованием исследовательского оборудования (пакетов прикладных программ)
	ПК-3.3 Способен оценивать точность полученных экспериментальных (численных) результатов
	ПК-3.4 Способен самостоятельно находить и осваивать новые информационные и программные ресурсы в области биоинженерии и биоинформатики
	ПК-3.5 Способен применять методы биоинженерии и биоинформатики для получения биологических объектов с целенаправленно измененными свойствами

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основные принципы и методы машинного обучения - понимание алгоритмов классификации, регрессии и кластеризации;
- ключевые концепции и технологии в области NLP - владение знаниями о токенизации, стемминге, лемматизации, POS-тэггинге и синтаксическом разборе;
- специфику и отличие разных типов медицинских данных;
- различные подходы и модели, используемые в NLP, архитектуру, отличия, преимущества и недостатки, в частности моделей BERT, GPT и LLaMa;
- этические и юридические аспекты работы с медицинскими данными, а именно конфиденциальность, защита данных и регулятивных требованиях в обработке медицинской информации.

уметь:

- применять Python для обработки текстовых данных - использовать библиотеки Python, такие как Pandas, NLTK, SpaCy, для предобработки, анализа и визуализации данных;
- разрабатывать и реализовывать модели NLP - построение и настройка моделей для автоматической классификации, извлечения информации и анализа тональности текстов;
- оценивать и интерпретировать результаты моделей машинного обучения. Требует анализа точности, полноты и других метрик качества моделей, адаптированных к медицинским задачам;
- внедрять разработанные модели и инструменты в реальные клинические и исследовательские процессы.

владеть:

- полученные знания и умения должны быть использованы для создания биомедицинских продуктов и технологий.

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. Основные алгоритмы и подходы в машинном обучении - классификация, регрессия, кластеризация. Основные техники и задачи NLP.		4		8
2	Введение в Python. Работа с различными типами данных: строками, списками, кортежами. Парсинг данных.		5		7
3	Python для NLP. Использование библиотек Python: Pandas, NumPy, Matplotlib, Библиотеки для NLP: NLTK, SpaCy.		3		8
4	Основы текстовой предобработки. Техники токенизации, стемминга, лемматизации, работа с текстовыми данными. POS-тэггинг и синтаксический разбор.		4		7
5	Датасеты и аннотация данных в медицинском NLP. Обзор популярных датасетов для медицинского NLP. Техники аннотации и валидации данных.		3		7

6	Введение в большие языковые модели. История создания больших языковых моделей. Развитие подходов к алгоритмам NLP: Bag-of-Words, Bag-of-Concepts, Bag-of-Narratives. Особенности медицинских моделей NLP.		4		8
7	Введение в трансформеры и модели BERT. Архитектура и принципы работы трансформеров. Вариации BERT для анализа медицинских данных. Основы обучения моделей, основанных на архитектуре BERT.		4		7
8	Переход от PLM (Pre-Trained Language Model) к LLM (Large Language Model). Модель GPT - создание и ее развитие. Примеры использования GPT в медицинских исследованиях.		3		8
9	Другие перспективные модели NLP в медицине (LLaMa, вариации GLM и др.).		3		8
10	Мультимодальные и мультиязычные модели NLP. Работа с различными языками и форматами данных. Примеры использования мультимодальных моделей в медицине. Подведение итогов по медицинским языковым моделям.		3		8
11	Визуализация данных. Инструменты визуализации для медицинских данных. Программы и библиотеки для визуализации. Кейсы визуализации в медицинском NLP.		4		9
12	Методы извлечения информации из медицинских текстов. Использование NLP для интерактомики и протеомики.		3		8
13	Протеогеномика и биоинформатика в NLP. Основы протеогеномики и ее приложения. Флаксомика и метаболомика в контексте NLP.		4		9
14	Валидация и оценка моделей NLP. Кросс-валидация и метрики оценки. Техники проверки и улучшения моделей. Интерпретация результатов моделей NLP.		3		8
15	Интеграция NLP в клинические и исследовательские процессы. Проекты по внедрению NLP в медицинские информационные системы. Разработка и тестирование клинических NLP-систем.		3		9
16	Этика и юридические аспекты в медицинском NLP. Защита данных и конфиденциальность. Этические вопросы работы с медицинскими данными.		4		8
17	Будущее медицинского NLP и искусственного интеллекта. Направления развития и новые возможности в медицинском NLP.		3		8

Итого часов		60		135
Подготовка к экзамену	30 час.			
Общая трудоёмкость	225 час., 5 зач.ед.			

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

Семестр: 1 (Осенний)

1. Введение в машинное обучение и NLP. Основные алгоритмы и подходы в машинном обучении - классификация, регрессия, кластеризация. Основные техники и задачи NLP.

История машинного обучения. Обучение с учителем. Обучение без учителя. Классификация: логистическая регрессия, деревья решений, случайный лес, SVM. Регрессия: линейная регрессия, полиномиальная регрессия, регрессия Ridge, регрессия Lasso. Кластеризация: K-means, DBSCAN, иерархическая кластеризация. Глубокое обучение: нейронные сети, обучение с подкреплением, сверточные нейронные сети, рекуррентные нейронные сети. Основы NLP: токенизация, морфологический анализ, синтаксический анализ, семантический анализ.

2. Введение в Python. Работа с различными типами данных: строками, списками, кортежами. Парсинг данных.

Синтаксис Python. Переменные и типы данных: числа, строки, булевы значения. Коллекции: списки, словари, множества, кортежи. Управляющие конструкции: условные операторы, циклы. Функции: определение функций, аргументы, возвращаемые значения. Обработка исключений. Работа с файлами: чтение, запись, обработка файлов в форматах TXT, CSV, JSON.

3. Python для NLP. Использование библиотек Python: Pandas, NumPy, Matplotlib, Библиотеки для NLP: NLTK, SpaCy.

Библиотеки для анализа данных: Pandas, NumPy. Введение в визуализацию данных: Matplotlib, Seaborn. Обработка текстов: регулярные выражения, NLTK, SpaCy. Извлечение именованных сущностей, анализ частот слов. Векторизация текста: Bag of Words, TF-IDF, Word2Vec, GloVe.

4. Основы текстовой предобработки. Техники токенизации, стемминга, лемматизации, работа с текстовыми данными. POS-тэггинг и синтаксический разбор.

Текстовая предобработка: токенизация, стемминг, лемматизация. POS-теггирование: определение частей речи. Синтаксический разбор: зависимостные и составные разборы. Семантический анализ: определение семантических ролей, анализ настроений. Работа с большими текстовыми корпусами: предобработка, нормализация, аннотация.

5. Датасеты и аннотация данных в медицинском NLP. Обзор популярных датасетов для медицинского NLP. Техники аннотации и валидации данных.

Медицинские датасеты: клинические записи, научные статьи, отчеты пациентов; различные типы медицинских данных: текстовые, числовые, графические, звуковые. Аннотация данных: методы ручной аннотации, использование инструментов автоматической аннотации. Валидация аннотированных данных: метрики качества, межаннотаторское согласие. Применение аннотированных данных в машинном обучении: обучение моделей, тестирование, оценка эффективности.

6. Введение в большие языковые модели. История создания больших языковых моделей. Развитие подходов к алгоритмам NLP: Bag-of-Words, Bag-of-Concepts, Bag-of-Narratives. Особенности медицинских моделей NLP.

История создания больших языковых моделей. Эволюция алгоритмов NLP: от ручных методов к машинному обучению. Подходы к алгоритмам NLP: Bag-of-Words, Bag-of-Concepts, Bag-of-Narratives. Специфика медицинских моделей NLP: терминология, контекстуальные особенности. Применение языковых моделей в медицинских исследованиях: анализ клинических записей, извлечение медицинской информации.

7. Введение в трансформеры и модели BERT. Архитектура и принципы работы трансформеров. Вариации BERT для анализа медицинских данных. Основы обучения моделей, основанных на архитектуре BERT.

Архитектура трансформера: механизм внимания, многослойные трансформер-блоки. Разработка и принципы работы BERT (Bidirectional Encoder Representations from Transformers). Вариации BERT для медицинских данных: BioBERT, Clinical BERT. Техники обучения моделей на основе трансформеров: fine-tuning, transfer learning. Анализ медицинских текстов с использованием BERT: классификация диагнозов, извлечение симптомов.

8. Переход от PLM (Pre-Trained Language Model) к LLM (Large Language Model). Модель GPT - создание и ее развитие. Примеры использования GPT в медицинских исследованиях.

Определение PLM (Pre-Trained Language Model) и LLM (Large Language Model). Разработка и эволюция модели GPT (Generative Pre-trained Transformer): от GPT-1 до GPT-4 (планируется также экскурс в GPT-5, которая должна выйти летом 2024 г.). Применение GPT в медицинских исследованиях: создание синтетических медицинских записей, автоматизация клинических резюме. Роль LLM в обработке медицинского языка: повышение точности, контекстуальное понимание.

Семестр: 2 (Весенний)

9. Другие перспективные модели NLP в медицине (LLaMa, вариации GLM и др.).

Развитие моделей NLP: LLaMa, GLM, CANINE, BigBird, ELMo. Особенности LLaMa: оптимизирована для различных языковых задач с минимальной настройкой. Применения GLM в медицинском NLP: адаптация к медицинской терминологии, улучшение интерпретируемости данных. CANINE: обработка символов для поддержки морфологического разнообразия. BigBird: обработка длинных последовательностей для комплексного анализа медицинских записей. ELMo: глубокое контекстуальное представление слов для повышения точности медицинского анализа.

10. Мультиязычные и мультимодальные модели NLP. Работа с различными языками и форматами данных. Примеры использования мультимодальных моделей в медицине. Подведение итогов по медицинским языковым моделям.

Мультиязычность в NLP: интеграция текста, изображений, звука. Мультимодальные модели NLP: работа с медицинскими текстами на различных языках. Примеры использования мультимодальных моделей в медицине: диагностика по изображениям и клиническим записям. Обзор достижений в медицинских языковых моделях: оценка эффективности, перспективы развития.

11. Визуализация данных. Инструменты визуализации для медицинских данных. Программы и библиотеки для визуализации. Кейсы визуализации в медицинском NLP.

Инструменты визуализации медицинских данных (интегрированных в различные датасеты) и результатов обработки алгоритмами NLP: Tableau, Power BI, библиотеки Python (Matplotlib, Seaborn, Plotly). Применение визуализации в медицинском NLP: анализ трендов, мониторинг распространения заболеваний. Кейсы использования визуализации: отображение результатов анализа больших данных, интерактивные дашборды для медицинских исследований.

12. Методы извлечения информации из медицинских текстов. Использование NLP для интерактомики и протеомики.

Использование NLP в интерактомике: анализ взаимодействия белков. Применение NLP в протеомике: идентификация белков, анализ белковых путей. Технологии извлечения информации: Named Entity Recognition, Relation Extraction. Примеры использования NLP для анализа научных публикаций, клинических протоколов.

13. Протеогеномика и биоинформатика в NLP. Основы протеогеномики и ее приложения. Флаксомика и метаболомика в контексте NLP.

Основы протеогеномики: интеграция геномных и протеомных данных. Применение протеогеномики в медицинском NLP: анализ молекулярных маркеров, биомаркеров. Флаксомика и метаболомика: анализ метаболических путей, предсказание биохимических взаимодействий. Роль NLP в биоинформатике: автоматизация анализа биологических данных, улучшение точности биомедицинских исследований.

14. Валидация и оценка моделей NLP. Кросс-валидация и метрики оценки. Техники проверки и улучшения моделей. Интерпретация результатов моделей NLP.

Кросс-валидация: Holdout, K-Fold, Stratified K-Fold. Метрики оценки: Accuracy, Precision, Recall, F1-Score, AUC-ROC. Техники проверки моделей: Confusion Matrix, ROC Curve Analysis. Интерпретация результатов: Feature Importance, LIME, SHAP для объяснения предсказаний модели.

15. Интеграция NLP в клинические и исследовательские процессы. Проекты по внедрению NLP в медицинские информационные системы. Разработка и тестирование клинических NLP-систем.

Проекты интеграции NLP: автоматизация клинического документирования, оптимизация поиска по медицинским базам данных. Разработка NLP-систем: анализ клинических записей, обработка и интерпретация медицинских изображений. Тестирование клинических NLP-систем: пилотные испытания, масштабирование и деплоймент.

16. Этика и юридические аспекты в медицинском NLP. Защита данных и конфиденциальность. Этические вопросы работы с медицинскими данными.

Защита данных: GDPR, HIPAA и их влияние на обработку медицинских данных. Конфиденциальность: анонимизация данных, безопасное хранение и передача. Этические вопросы: согласие на обработку данных, риски неправильной интерпретации результатов. Юридические аспекты: ответственность за ошибки AI, правовые рамки использования медицинского AI (в РФ и других странах).

17. Будущее медицинского NLP и искусственного интеллекта. Направления развития и новые возможности в медицинском NLP.

Направления развития: расширение возможностей автоматизации, углубление понимания комплексных медицинских сценариев. Расширение возможностей мультимодальных моделей, построение новых мультимодальных моделей. Улучшенная обработка числовых значений. Усовершенствование медицинских роботов при помощи NLP. Квантовые вычисления для NLP, интеграция геномных данных для персонализированной медицины.

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

1. Введение в методы машинного обучения с подкреплением, учебное пособие /А. И. Панов; Министерство науки и высшего образования Российской Федерации ; Московский физико-технический институт (национальный исследовательский университет). Москва, МФТИ, 2019
2. Python и машинное обучение [Текст], / С. Рашка; пер. с англ. А. В. Логунова, М., ДМК Пресс, 2017

Дополнительная литература

Рекомендуемая литература для самостоятельного изучения:

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>.

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)

Успешное освоение курса требует напряжённой самостоятельной работы студента. В программе курса приведено минимально необходимое время для работы студента над темой.

Самостоятельная работа включает в себя:

- проработку учебного материала (по конспектам лекций, учебной и научной литературе);
- подготовку ответов на вопросы, предназначенные для самостоятельного изучения, доказательство отдельных утверждений, свойств.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Биотехнология
профиль подготовки: Системная и синтетическая биология
Физтех-школа Биологической и Медицинской Физики
кафедра системной и синтетической биологии
курс: 1
квалификация: магистр

Семестры, формы промежуточной аттестации:
1 (осенний) - Зачет
2 (весенний) - Экзамен

Разработчик: К.Д. Балбек

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
ОПК-2 Имеет представление об актуальных проблемах науки и техники в области своей профессиональной деятельности, способен на научном языке формулировать профессиональные задачи	ОПК-2.1 Имеет представление о современном состоянии исследований в рамках тематической области своей профессиональной деятельности
	ОПК-2.2 Способен оценивать актуальность исследований в области своей профессиональной деятельности и их практическую значимость
	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
ОПК-3 Способен выбирать и (или) разрабатывать подходы к решению типовых и новых задач в области профессиональной деятельности, учитывая особенности и ограничения различных методов решения	ОПК-3.1 Способен анализировать задачу, планировать пути решения, предлагать и комбинировать способы решения
	ОПК-3.2 Способен использовать исследовательские методы при решении новых задач, применяя знания в различных областях науки (техники)
	ОПК-3.3 Владеет аналитическими и вычислительными методами решения, понимает и учитывает на практике границы применимости получаемых решений
	ОПК-3.4 Способен к профессиональной эксплуатации и модернизации современного технологического оборудования для осуществления биотехнологических процессов
	ОПК-3.5 Владеет навыками проектирования новых биотехнологических решений для поставленных научно-технических и технологических задач
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.2 Способен использовать специализированные знания фундаментальных разделов математики, физики, химии и биологии для постановки и решения научно-исследовательских задач в области биоинженерии и биоинформатики
	ПК-1.3 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.4 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты

	ПК-1.5 Способен создавать программные средства и базы данных, используемые в биоинженерии и биоинформатике
ПК-3 Способен профессионально работать с исследовательским и испытательным оборудованием (приборами и установками, специализированными пакетами прикладных программ) в избранной предметной области	ПК-3.1 Понимает принципы работы используемого оборудования (специализированных пакетов прикладных программ)
	ПК-3.2 Способен проводить эксперимент (моделирование) с использованием исследовательского оборудования (пакетов прикладных программ)
	ПК-3.3 Способен оценивать точность полученных экспериментальных (численных) результатов
	ПК-3.4 Способен самостоятельно находить и осваивать новые информационные и программные ресурсы в области биоинженерии и биоинформатики
	ПК-3.5 Способен применять методы биоинженерии и биоинформатики для получения биологических объектов с целенаправленно измененными свойствами

2. Показатели оценивания компетенций

В результате изучения дисциплины «Основы алгоритмов обработки естественного языка в индустрии здравоохранения» обучающийся должен:

знать:

- основные принципы и методы машинного обучения - понимание алгоритмов классификации, регрессии и кластеризации;
- ключевые концепции и технологии в области NLP - владение знаниями о токенизации, стемминге, лемматизации, POS-тэггинге и синтаксическом разборе;
- специфику и отличие разных типов медицинских данных;
- различные подходы и модели, используемые в NLP, архитектуру, отличия, преимущества и недостатки, в частности моделей BERT, GPT и LLaMa;
- этические и юридические аспекты работы с медицинскими данными, а именно конфиденциальность, защита данных и регулятивных требованиях в обработке медицинской информации.

уметь:

- применять Python для обработки текстовых данных - использовать библиотеки Python, такие как Pandas, NLTK, SpaCy, для предобработки, анализа и визуализации данных;
- разрабатывать и реализовывать модели NLP - построение и настройка моделей для автоматической классификации, извлечения информации и анализа тональности текстов;
- оценивать и интерпретировать результаты моделей машинного обучения. Требует анализа точности, полноты и других метрик качества моделей, адаптированных к медицинским задачам;
- внедрять разработанные модели и инструменты в реальные клинические и исследовательские процессы.

владеть:

- полученные знания и умения должны быть использованы для создания биомедицинских продуктов и технологий.

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

Во время текущего контроля студент должен уметь ответить на следующие вопросы:

1. Ключевые библиотеки Python для NLP, способы применения.
2. Техники предобработки текста, используемые в NLP.
3. Основные типы медицинских данных; основные датасеты для разработки медицинского ИИ.
4. Как развивались подходы к алгоритмам NLP? Отличия Bag-Of-Words, Bag-of-Concepts, Bag-of-Narratives.

5. Устройство архитектуры трансформеров. Устройство архитектуры моделей, основанных на BERT.
6. Методы деперсонализации медицинских данных.

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

Типовые вопросы для экзамена:

1. Различия между supervised, unsupervised и reinforcement learning в контексте NLP.
2. Мультиязычные NLP-системы в медицине - преимущества и недостатки.
3. Методы определения семантической близости в текстах.
4. Принципы работы и преимущества использования мультимодальных моделей NLP.
5. Подходы для уменьшения переобучения в глубоких нейронных сетях.
6. Примеры использования NLP для обработки естественного языка в клинических записях (EHR).
7. Типы наиболее эффективных нейронных сетей для задач по классификации текстов.
8. Особенности модели BioBERT по сравнению с обычной моделью BERT для медицинских данных.
9. Механизм внимания в трансформерах (Attention).
10. Процесс обучения нейронной сети на примере сети LSTM для обработки временных рядов.

Примеры билетов для экзамена:

Билет 1:

1. Защита курсового проекта.
2. Устройство, преимущества и недостатки модели LLaMa для улучшения понимания медицинских текстов.

Билет 2:

1. Защита курсового проекта.
2. Техники машинного обучения, применяемые для определения посттрансляционных модификаций в протеомике.

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.

Оценка «зачтено» выставляется студенту, если он показал 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 «не зачтено» 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.

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

При проведении зачета и экзамена обучающемуся предоставляется 45 минут на подготовку. Опрос обучающегося не должен превышать одного астрономического часа.