

**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:	Анализ естественного языка
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: 3 (осенний) - Дифференцированный зачет

Аудиторных часов: 28 всего, в том числе:

лекции: 2 час.

семинары: 26 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 62 час.

Всего часов: 90, всего зач. ед.: 2

Программу составил: К.А. Лапин, методист

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

Аннотация

В курсе рассматриваются задачи, которые требуют обработки текстов на естественных языках, в первую очередь русском и английском. Список задач включает в себя классификацию текстов, определение тональности, автоматическое реферирование, машинный перевод, многие другие задачи более низкого уровня. Из подходов к решению задач рассматриваются лингвистические подходы, статистические и подходы, использующие глубокое обучение. Курс предполагает решение практических заданий с помощью библиотек и ресурсов для обработки естественных языков.

1. Цели и задачи

Цель дисциплины

- приобретение навыков решения практических заданий с помощью библиотек и ресурсов для обработки естественных языков.

Задачи дисциплины

- сформировать навыки обработки неструктурированного текста на естественном языке;
- сформировать навыки работы с современными математическими методами и компьютерными алгоритмами для решения этих задач.

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-2 Способен управлять проектом на всех этапах его жизненного цикла	УК-2.1 Формулирует в рамках обозначенной проблемы цель, задачи, актуальность, значимость (научную, практическую, методическую и иную в зависимости от типа проекта), ожидаемые результаты и возможные сферы их применения
	УК-2.2 Способен прогнозировать результат деятельности и планировать последовательность шагов для достижения данного результата. Формирует план-график реализации проекта в целом и план контроля его выполнения
	УК-2.3 Способен организовать и координировать работу участников проекта, обеспечивать работу команды необходимыми ресурсами
УК-4 Способен применять современные коммуникативные технологии, в том числе на иностранном(ых) языке(ах), для академического и профессионального взаимодействия	УК-4.1 Способен вести обмен деловой информацией в устной и письменной формах на государственном языке Российской Федерации и не менее чем на одном иностранном языке
	УК-4.2 Владеет навыками, необходимыми для написания, письменного перевода и редактирования различных академических текстов (рефератов, эссе, обзоров, статей и т.д.)

взаимодействия	УК-4.3 Способен представлять результаты академической и профессиональной деятельности на различных научных мероприятиях, включая международные
УК-6 Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки	УК-6.1 Умеет решать задачи собственного личностного и профессионального развития, определять и реализовывать приоритеты совершенствования собственной деятельности
	УК-6.2 Оценивает свою деятельность, соотносит цели, способы и средства выполнения деятельности с её результатами
ОПК-1 Способен решать актуальные задачи фундаментальной и прикладной математики	ОПК-1.1 Знает и способен использовать в профессиональной деятельности фундаментальные научные знания и новые научные принципы и методы исследований в области прикладной математики и информатики
	ОПК-1.2 Способен обобщать и критически оценивать опыт и результаты научных исследований в области профессиональной деятельности
	ОПК-1.3 Понимает междисциплинарные связи в области прикладной математики и информатики и способен их применять при решении задач профессиональной деятельности
ОПК-6 Способен выбирать и (или) разрабатывать подходы к решению типовых и новых задач в области информатики и вычислительной техники, учитывая особенности и ограничения различных методов решения	ОПК-6.1 Способен анализировать задачу, планировать пути решения, предлагать и комбинировать способы решения
	ОПК-6.2 Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем
	ОПК-6.5 Способен самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте
ПК-10 Способен применять методы планирования исследований и экспериментов при выполнении проектов и заданий в избранной предметной области	ПК-10.2 Умеет применять теоретические знания к построению программ исследований и экспериментов при выполнении конкретных проектов и заданий
	ПК-10.3 Владеет методами планирования исследований и экспериментов в избранной предметной области

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- методы анализа проблемных ситуаций как систем;
- методы обмена деловой информацией в устной и письменной формах на государственном языке Российской Федерации и не менее чем на одном иностранном языке;
- о современном состоянии исследований в рамках тематической области своей профессиональной деятельности;
- методы анализа задач, а также методы планирования их решения;
- методы и средства разработки программного обеспечения, методы управления проектами разработки программного обеспечения, способы организации проектных данных, нормативно-технические документы (стандарты и регламенты) по разработке программных средств и проектов.

уметь:

- осуществлять поиск вариантов решения поставленной проблемной ситуации на основе доступных источников информации;
- представлять результаты академической и профессиональной деятельности на различных научных мероприятиях, включая международные;
- оценивать актуальность исследований в области информатики и вычислительной техники и их практическую значимость;
- анализировать и вычислять решения задач, понимать и учитывать на практике границы применимости получаемых решений;
- выбирать средства разработки, оценивать сложность проектов, планировать ресурсы, контролировать сроки выполнения и оценивать качество полученного результата.

владеть:

- навыками разработки стратегий достижения поставленной цели как последовательности шагов, предвидя результат каждого из них и оценивая их влияние на внешнее окружение планируемой деятельности и на взаимоотношения участников этой деятельности;
- навыками использования современных средств информационно-коммуникационных технологий для академического и профессионального взаимодействия;
- профессиональной терминологией, используемой в современной научно-технической литературе, обладает навыками устного и письменного изложения результатов научной деятельности в рамках профессиональной коммуникации;
- навыками самостоятельного приобретения, развития и применения математических, естественнонаучных, социально-экономических и профессиональных знаний для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте;
- методами разработки технического задания, составления планов, распределения задач, тестирования и оценки качества программных средств.

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	Основные NLP задачи и их использование в бизнесе		3		10
2	Векторные представления текстов	2			12
3	Рекуррентные нейронные сети		3		10
4	Transfer learning in NLP		7		10
5	Задача Token Classification		7		10
6	Задача Question answering		6		10
Итого часов		2	26		62
Подготовка к экзамену		0 час.			
Общая трудоёмкость		90 час., 2 зач.ед.			

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

Семестр: 3 (Осенний)

1. Основные NLP задачи и их использование в бизнесе

Классификация текстов. Классификация отдельных слов в тексте. Машинный перевод. Поиск ответа на вопрос. Задачи генерации текста. Этические проблемы использования NLP моделей.

2. Векторные представления текстов

Отображение текста в векторное пространство. Статистические методы. Word2Vec.

3. Рекуррентные нейронные сети

Рекуррентные нейронные сети. RNN для задачи текстовой классификации.

4. Transfer learning in NLP

Архитектура Transformer. State-of-the-art-модели. Transfer learning.

5. Задача Token Classification

Задача Token Classification. Задача NER(Named Entity Recognition). Токенайзер BERT на практике.

6. Задача Question answering

Процесс решения задачи поиска ответа на вопрос в контексте (question answering). Способы оценки качества решения. Подходы к решению такого рода задач с использованием современных архитектур

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

Занятия проходят с использованием дистанционных образовательных технологий.

6. Перечень рекомендуемой литературы

Основная литература

1. Анализ алгоритмов [Текст] : вводный курс / Дж. Макконелл ; пер. с англ. С. К. Ландо .— М. : Техносфера, 2002 .— 304 с.
2. Анализ алгоритмов. Активный обучающий подход [Текст] : учеб. пособие для вузов / Дж. Макконелл ; пер. с англ. С. А. Кулешова ; под ред. С. К. Ландо .— 3-е изд., доп. — М. : Техносфера, 2013 .— 416 с.

Дополнительная литература

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

Журнал «Искусственный интеллект и принятие решений», Труды конференции Диалог
<http://www.dialog-21.ru/>, A Digital Archive of
Research Papers in Computational Linguistics <http://aclweb.org/anthology/>.

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

Студенту для занятий потребуются:

1. Google Drive для доступа к материалам курса
2. Zoom

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

Студенту рекомендуется внимательно слушать лектора, следить за тем, что написано на доске или представлено на слайдах презентации, анализировать получаемую им информацию. В случае, если материал лекции непонятен, следует задать вопрос в отведенное для вопросов время. Студенту также рекомендуется конспектировать материал лекции в тетради, что улучшает запоминание.

При выполнении практических работ студенту рекомендуется внимательно анализировать поставленную задачу, уделяя особое внимание критериям оценки точности решения задачи. Особенное внимание следует уделять методологическим аспектам решения задач.

При ведении самостоятельной работы студенту рекомендуется внимательно подходить к изучению научных статей, обращать внимание на значимость полученного результата, на требования к обучающей выборке, на скорость работы предлагаемых алгоритмов, на результаты их сравнения с существующими. В случае, если изучаемый материал понятен не до конца, рекомендуется обращение к дополнительной литературе.

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: 3 (осенний) - Дифференцированный зачет	
Разработчик:	К.А. Лапин, 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-2 Способен управлять проектом на всех этапах его жизненного цикла	УК-2.1 Формулирует в рамках обозначенной проблемы цель, задачи, актуальность, значимость (научную, практическую, методическую и иную в зависимости от типа проекта), ожидаемые результаты и возможные сферы их применения
	УК-2.2 Способен прогнозировать результат деятельности и планировать последовательность шагов для достижения данного результата. Формирует план-график реализации проекта в целом и план контроля его выполнения
	УК-2.3 Способен организовать и координировать работу участников проекта, обеспечивать работу команды необходимыми ресурсами
УК-4 Способен применять современные коммуникативные технологии, в том числе на иностранном(ых) языке(ах), для академического и профессионального взаимодействия	УК-4.1 Способен вести обмен деловой информацией в устной и письменной формах на государственном языке Российской Федерации и не менее чем на одном иностранном языке
	УК-4.2 Владеет навыками, необходимыми для написания, письменного перевода и редактирования различных академических текстов (рефератов, эссе, обзоров, статей и т.д.)
	УК-4.3 Способен представлять результаты академической и профессиональной деятельности на различных научных мероприятиях, включая международные
УК-6 Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки	УК-6.1 Умеет решать задачи собственного личностного и профессионального развития, определять и реализовывать приоритеты совершенствования собственной деятельности
	УК-6.2 Оценивает свою деятельность, соотносит цели, способы и средства выполнения деятельности с её результатами
ОПК-1 Способен решать актуальные задачи фундаментальной и прикладной математики	ОПК-1.1 Знает и способен использовать в профессиональной деятельности фундаментальные научные знания и новые научные принципы и методы исследований в области прикладной математики и информатики
	ОПК-1.2 Способен обобщать и критически оценивать опыт и результаты научных исследований в области профессиональной деятельности
	ОПК-1.3 Понимает междисциплинарные связи в области прикладной математики и информатики и способен их применять при решении задач профессиональной деятельности

ОПК-6 Способен выбирать и (или) разрабатывать подходы к решению типовых и новых задач в области информатики и вычислительной техники, учитывая особенности и ограничения различных методов решения	ОПК-6.1 Способен анализировать задачу, планировать пути решения, предлагать и комбинировать способы решения
	ОПК-6.2 Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем
	ОПК-6.5 Способен самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте
ПК-10 Способен применять методы планирования исследований и экспериментов при выполнении проектов и заданий в избранной предметной области	ПК-10.2 Умеет применять теоретические знания к построению программ исследований и экспериментов при выполнении конкретных проектов и заданий
	ПК-10.3 Владеет методами планирования исследований и экспериментов в избранной предметной области

2. Показатели оценивания компетенций

В результате изучения дисциплины «Анализ естественного языка» обучающийся должен:

знать:

- методы анализа проблемных ситуаций как систем;
- методы обмена деловой информацией в устной и письменной формах на государственном языке Российской Федерации и не менее чем на одном иностранном языке;
- о современном состоянии исследований в рамках тематической области своей профессиональной деятельности;
- методы анализа задач, а также методы планирования их решения;
- методы и средства разработки программного обеспечения, методы управления проектами разработки программного обеспечения, способы организации проектных данных, нормативно-технические документы (стандарты и регламенты) по разработке программных средств и проектов.

уметь:

- осуществлять поиск вариантов решения поставленной проблемной ситуации на основе доступных источников информации;
- представлять результаты академической и профессиональной деятельности на различных научных мероприятиях, включая международные;
- оценивать актуальность исследований в области информатики и вычислительной техники и их практическую значимость;
- анализировать и вычислять решения задач, понимать и учитывать на практике границы применимости получаемых решений;
- выбирать средства разработки, оценивать сложность проектов, планировать ресурсы, контролировать сроки выполнения и оценивать качество полученного результата.

владеть:

- навыками разработки стратегий достижения поставленной цели как последовательности шагов, предвидя результат каждого из них и оценивая их влияние на внешнее окружение планируемой деятельности и на взаимоотношения участников этой деятельности;
- навыками использования современных средств информационно-коммуникационных технологий для академического и профессионального взаимодействия;
- профессиональной терминологией, используемой в современной научно-технической литературе, обладает навыками устного и письменного изложения результатов научной деятельности в рамках профессиональной коммуникации;
- навыками самостоятельного приобретения, развития и применения математических, естественнонаучных, социально-экономических и профессиональных знаний для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте;
- методами разработки технического задания, составления планов, распределения задач, тестирования и оценки качества программных средств.

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

Текущий контроль успеваемости по дисциплине осуществляется решения практических задач, входящих в состав курса.

Примеры практических заданий:

1. Основные NLP задачи и их использование в бизнесе

Используя класс `pipeline` из библиотеки `transformers` и подходящие для этого модели, найдите в тексте модуля ответы на перечисленные вопросы.

Выполните это задание в ноутбуке (можно воспользоваться шаблоном) в котором приведете инференс выбранной вами модели для `question answering` на русском языке для выбранных вами абзацев текста модуля, и кратко проанализируйте полученные ответы.

- Какая задача может быть развитием задачи классификации текстов?
- Что послужило основным толчком для существенного улучшения качества решения NLP задач в последние несколько лет?
- Что служит входными данными для задачи поиска ответа на вопрос в заданном контексте?
- Какие NLP задачи успешно решались с использованием классических статистических методов?

2. Векторные представления текстов

Ответьте на этот вопрос без реализации, описав свои идеи.

Подумайте, как лучше всего стоит обрабатывать случаи, когда мы хотим получить векторы для устойчивых фраз/словосочетаний/терминов, которые состоят из нескольких слов. Например, «Нижний Новгород», «Великий Устюг» и т.д.

2. Подготовьте небольшой анализ понравившейся вам работы из следующего списка:

- Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings
- Gender Bias in Contextualized Word Embeddings
- Analogies Explained: Towards Understanding Word Embeddings
- The (Too Many) Problems of Analogical Reasoning with Word Vectors
- Frustratingly Easy Meta-Embedding – Computing Meta-Embeddings by Averaging Source Word Embeddings
- Understanding the Origins of Bias in Word Embeddings

3. Рекуррентные нейронные сети

1. Ответьте на следующий вопрос, описав свои идеи без реализации.

Подумайте, как можно попробовать решить проблему `exploding gradients`, помимо `gradient clipping`? Как можно подойти к решению проблемы `vanishing gradients`?

2. Подготовьте небольшой анализ понравившейся вам работы из следующего списка:

- Improving Tree-LSTM with Tree Attention
- Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach
- On the Effectiveness of Low-Rank Matrix Factorization for LSTM Model Compression
- Sentence-State LSTM for Text Representation

4. Transfer learning in NLP

Решить задачу классификации текстов, используя трансформер и сравнить полученные оценки качества. Для примера обучения трансформера под задачу классификации можете воспользоваться приведенным в модуле ноутбуком с решением задачи классификации на `Ag-news`.

5. Задача Token Classification

1. Решить задачу NER для предоставленного датасета, используя любые доступные вам средства. Модель должна обучаться на файле `train.txt`, валидироваться на файле `dev.txt` и её качество необходимо оценить на файле `test.txt`.

Для достижения наилучшего результата уделите внимание подбору гиперпараметров как в плане архитектуры, так и в плане гиперпараметров обучения модели.

2. Задача Question answering

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

Примерные вопросы к дифференцированному зачету:

1. Рекуррентные нейронные сети. RNN для задачи текстовой классификации.
2. Классификация текстов.
3. Классификация отдельных слов в тексте.
4. Машинный перевод. Поиск ответа на вопрос.
5. Задачи генерации текста.
6. Этические проблемы использования NLP моделей.
7. Лингвистические приложения методов классификации.
8. Классификация элементов последовательности символов и слов
9. Решить задачу QA для датасета SberQuad, используя любые доступные вам средства.
10. Для датасета Ag-news произведите fine-tuning модели BERT и сравните производительность на тестовом множестве с производительностью нескольких дистиллированных версий BERT (кроме классического DistilBERT рекомендуется взять ещё несколько), а также произведите несколькими способами квантизацию модели и оцените полученные метрики качества.

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 осуществляется в виде дифференцированного зачета на lms.

Время отведенное на дифференцированный зачет: 4 академических часа.