

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

УТВЕРЖДЕНО

**Директор физтех-школы
прикладной математики и
информатики**

А.М. Райгородский

Рабочая программа дисциплины (модуля)

по дисциплине:	NLP. Обработка естественного языка
по направлению:	Прикладная математика и информатика
профиль подготовки:	Проектирование и разработка комплексных бизнес-приложений Физтех-школа Прикладной Математики и Информатики кафедра машинного обучения и цифровой гуманитаристики
курс:	4
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 7 (осенний) - Дифференцированный зачет

Аудиторных часов: 60 всего, в том числе:

лекции: 30 час.

семинары: 30 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 75 час.

Всего часов: 135, всего зач. ед.: 3

Программу составил: Р.Г. Нейчев, старший преподаватель

Программа обсуждена на заседании кафедры машинного обучения и цифровой гуманитаристики 14.03.2023

Аннотация

Современные подходы в различных областях искусственного интеллекта основаны на методах глубокого обучения (например, в компьютерном зрении, обработке естественного языка, обучении с подкреплением и т. Д.). Глубокие нейронные архитектуры демонстрируют большой потенциал и обещают даже лучшие результаты, поэтому сейчас определенно самое лучшее. время исследовать эту область.

В этом курсе мы начнем с основ и быстро погрузимся в последние результаты в области обработки естественного языка, уделяя особое внимание новым подходам и прикладным методам. Этот курс имеет тенденцию развивать как практические навыки, так и теоретическую базу, чтобы предоставить студентам глубокие теоретические знания и способность работать самостоятельно над проектами НЛП.

1. Цели и задачи

Цель дисциплины

- Познакомьтесь с классическими и новыми техниками в области НЛП.
- Получите практический опыт решения задач обработки естественного языка.
- Развивать навыки применения моделей НЛП к реальным данным.

Задачи дисциплины

- Постановка задачи обработки естественного языка и возможность разработать общий конвейер решения.
- Выберите подходящий подход и модель для конкретной проблемы.
- Существенный опыт работы с фреймворком PyTorch и Python

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области
	ОПК-2.3 Знает основные требования информационной безопасности
ОПК-3 Способен составлять и оформлять научные и (или) технические (технологические, инновационные) отчеты (публикации, проекты)	ОПК-3.1 Знает основные правила оформления научных публикаций и научно-технической документации, в том числе с использованием прикладного программного обеспечения
	ОПК-3.2 Владеет на практике методологией составления научно-технических отчетов (проектов)
	ОПК-3.3 Владеет методами визуального и графического представления результатов научной (научно-технической, инновационной технологической) деятельности в виде отчетов, научных публикаций
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации
	ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива
	ПК-2.3 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и участия в конференциях

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны знать:

- постановку задач морфологического, синтаксического анализа;
- методы решения этих задач.

уметь:

- формулировать задачи классификации текстов, предложений или их элементов для выделения структурированной информации;
- реализовывать подходящий алгоритм классификации текстов;
- решать задачи выделения ключевых слов и определения тональности.

владеть:

- основными программными системами для выделения скрытых тем и снижения размерности векторных моделей.

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	Классические подходы к векторизации текста: BoW, TF-IDF.	7	6		15
2	Взрыв в глубоких нейронных сетях.	5	6		15
3	Поиск луча	7	6		15
4	Внимание к архитектуре кодер-декодер.	5	6		15
5	Обзор семейства GPT.	6	6		15
Итого часов		30	30		75
Подготовка к экзамену		0 час.			
Общая трудоёмкость		135 час., 3 зач.ед.			

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

Семестр: 7 (Осенний)

1. Классические подходы к векторизации текста: BoW, TF-IDF.

Текстовые сочетания Вложения слов; word2vec и GLoVe языковые модели подход seq2seq

2. Взрыв в глубоких нейронных сетях.

Сверточные нейронные сети в НЛП. CNN для обработки текста
Машинный перевод и нейронный машинный перевод.

3. Поиск луча

Измерение качества сгенерированного текста. BLEU / Оценка недоумения. Механизм внимания. Механизм самовнимания

4. Внимание к архитектуре кодер-декодер.

Обзор архитектуры трансформатора. Предварительная подготовка по НЛП. Контекстные вложения. ELMo. Обзор BERT.

5. Обзор семейства GPT.

Системы ответов на вопросы и знания. Двухнаправленный поток внимания (BiDAF) Анализ настроений POS-теги, парсинг зависимостей

Тематическое моделирование (PLSA. LDA) Техники RL в НЛП. Обучение самокритичной последовательности.

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

Стандартная аудитория

6.Перечень рекомендуемой литературы

Основная литература

Литература кафедры:

1. Машинное обучение [Текст]/Х. Бринк, Дж. Ричардс, М. Феверолф, Real-World Machine Learning, -СПб., Питер, 2017

2. Python и машинное обучение [Текст] = Python Machine Learning : крайне необходимое издание по новейшей предсказательной аналитике для более глубокого понимания методологии машинного обучения / С. Рашка; пер. с англ. А. В. Логунова. — М. : ДМК Пресс, 2017. — 418 с.: ил. - Предм. указ.: с. 408-417. - 200 экз. - ISBN 978-5-97060-409-0 (в пер.) .— Полный текст (Доступ из сети МФТИ / Удаленный доступ).

Дополнительная литература

Литература кафедры:

1. Математические основы машинного обучения и прогнозирования [Текст] / В. В. Вьюгин ; Моск. физ.-техн. ин-т (гос. ун-т), Лаб. структурных методов анализа данных в предсказательном моделировании (ПреМоЛаб), Ин-т проблем передачи информации им. А. А. Харкевича РАН - М.МЦНМО,2013

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

<http://dm.fizteh.ru/>

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

Мультимедийные технологии можно использовать на лекциях и практических занятиях, в том числе на презентациях.

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, понятия, аксиомы.

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:

- чтение и конспектирование рекомендованной литературы;
- проработку учебного материала (по учебной и научной литературе), подготовку ответов на вопросы, предназначенных для самостоятельного изучения, доказательство отдельных утверждений, свойств;
- подготовку к дифференцированному зачету.

Руководство и контроль за самостоятельной работой студента осуществляется в форме индивидуальных консультаций.

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

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению:	Прикладная математика и информатика
профиль подготовки:	Проектирование и разработка комплексных бизнес-приложений Физтех-школа Прикладной Математики и Информатики кафедра машинного обучения и цифровой гуманитаристики
курс:	<u>4</u>
квалификация:	бакалавр
Семестр, формы промежуточной аттестации: 7 (осенний) - Дифференцированный зачет	
Разработчик:	Р.Г. Нейчев, старший преподаватель

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области
	ОПК-2.3 Знает основные требования информационной безопасности
ОПК-3 Способен составлять и оформлять научные и (или) технические (технологические, инновационные) отчеты (публикации, проекты)	ОПК-3.1 Знает основные правила оформления научных публикаций и научно-технической документации, в том числе с использованием прикладного программного обеспечения
	ОПК-3.2 Владеет на практике методологией составления научно-технических отчетов (проектов)
	ОПК-3.3 Владеет методами визуального и графического представления результатов научной (научно-технической, инновационной технологической) деятельности в виде отчетов, научных публикаций
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации
	ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива
	ПК-2.3 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и участия в конференциях

2. Показатели оценивания компетенций

В результате изучения дисциплины «NLP. Обработка естественного языка» обучающийся должен:

знать:

- постановку задач морфологического, синтаксического анализа;
- методы решения этих задач.

уметь:

- формулировать задачи классификации текстов, предложений или их элементов для выделения структурированной информации;
- реализовывать подходящий алгоритм классификации текстов;
- решать задачи выделения ключевых слов и определения тональности.

владеть:

- основными программными системами для выделения скрытых тем и снижения размерности векторных моделей.

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

1. Представления слов: основные подходы (BoW, TF-IDF).
2. Вложения слов (word2vec: линейность, скийп-грамма, отрицательная выборка, ключевые идеи)
3. RNN в обработке текста. Контекст, память
4. CNN в обработке текста. Отношение к подходу n-грамматики.
5. Механизм внимания.
6. Механизм самовнимания.

7. Контекстуализированные вложения.
8. Архитектура преобразователя: основные детали структуры кодировщика и декодера.
9. Архитектура BERT. Основные идеи (маскировка, предварительное обучение по многим задачам)
10. Метрики машинного перевода, функции качества.
11. Предварительная подготовка по проблемам НЛП.
12. Предвзятость при формировании языка
13. Вопросно-ответные системы: ключевые понятия
14. Подход к обучению самокритичной последовательности.

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

Вопросы к экзамену

1. Докажите, что если m , n - два взаимно простых целых числа разной четности, то числа $m^2 - n^2$ и $2mn$ также взаимно просты.
2. Напишите и докажите общую формулу для количества различных представлений данного целого числа n в виде суммы двух квадратов. (Представители, которые не получены друг от друга путем изменения знаков и порядка слов, считаются разными.)
3. На основе полученной формулы выведите нижнюю границу максимального числа равных расстояний между заданными n точками на плоскости, используя правильную прямоугольную решетку.
4. Постройте правильный пятиугольник с помощью циркуля и линейки.
5. Постройте правильный 15-угольник, используя циркуль и линейку.
6. Вам дается один сегмент. Требуется построить с помощью циркуля и линейки отрезок длины x , удовлетворяющий уравнению
7. Основываясь на предыдущем задании, докажите, что правильный семиугольник нельзя построить с помощью циркуля и линейки.
8. Докажите, что трисекция угла невозможна.
9. Опишите все возможные комбинации количества черных и белых шаров в урне для голосования, чтобы при случайном вылове двух шаров в выборке без возврата, вероятность вылова двух белых шаров составляла точно 0,5.
10. Рассмотрим соотношение сторон a , b , c треугольника, в котором треугольник с вершинами в основании биссектрис равнобедренный. Предполагая, что стороны, сходящиеся на стороне с большого треугольника, равны, сведем это соотношение к следующему
11. Далее мы рассматриваем куб, определяемый первым из трех уравнений (отказ от требования, чтобы a , b , c были сторонами треугольника). Покажите, что полученный куб неразложим, то есть определяющий его многочлен не учитывается.
12. В дополнение к этому, покажите, что наш куб неособен, то есть на его проективизации нет ни одной точки, в которой каждое направление касалось бы (или того же самого, в котором все три первые частные производные многочлена, определяющего его, вырождаются.).

Примеры экзаменационных билетов

Билет №1

1. Напишите и докажите общую формулу для количества различных представлений данного целого числа n в виде суммы двух квадратов.
2. Докажите, что трисекция угла невозможна.

Билет №2

1. Рассмотрим соотношение сторон a , b , c треугольника, в котором треугольник с вершинами в основании биссектрис равнобедренный.
2. Опишите всевозможные комбинации чисел черных и белых шаров в урне для голосования так, чтобы, если два шара случайно выловлены в выборке и не вернулись, вероятность вылова двух белых шаров была ровно 0,5.

Критерии оценивания

Оценка «отлично (10)» выставляется студенту, который проявил всестороннее, систематическое и глубокое знание материала образовательной программы, самостоятельно выполнил все задачи, предусмотренные программой, глубоко изучил основную и дополнительную литературу, рекомендованную программой. , активно работает в классе и понимает основные научные концепции по изучаемой дисциплине, проявил творческий подход и научный подход в понимании и представлении материала образовательной программы, ответ на который характеризуется использованием богатых и адекватных терминов, а также последовательным и логичным изложением материала;

Оценка «отлично (9)» дается студенту, который продемонстрировал всестороннее систематическое знание материала образовательной программы, самостоятельно выполнил все задачи, предусмотренные программой, глубоко усвоил основную литературу и знаком с рекомендуемой дополнительной литературой. по программе, активно проработал на занятиях, показал системность знаний по дисциплине, достаточную для дальнейшего изучения, а также умение самостоятельно расширять ее, ответ которой отличается точностью используемых терминов, а изложение материала в нем последовательное и логичное;

Оценка «отлично (8)» выставляется студенту, который проявил полное знание материала образовательной программы, не допускает существенных неточностей в своем ответе, самостоятельно выполнил все задания, предусмотренные программой, изучил основную литературу, рекомендованную учебной программой. программа, активно проработанная на занятиях, показала системность его знаний по дисциплине, достаточных для дальнейшего изучения, а также способность самостоятельно их расширять;

Оценка «хорошо (7)» выставляется студенту, который проявил достаточно полное знание материала образовательной программы, не допускает существенных неточностей в ответе, самостоятельно выполнил все задания, предусмотренные программой, изучил основную рекомендованную литературу по программе, активно работал на занятиях, проявил системность своих знаний по дисциплине, достаточных для дальнейшего изучения, а также способность самостоятельно их усиливать;

Оценка «хорошо (6)» выставляется студенту, который проявил достаточно полное знание материала образовательной программы, не допускает существенных неточностей в своем ответе, самостоятельно выполнил основные задачи, предусмотренные программой, изучил основную литературу. рекомендован программой, показал систематичность своих знаний по дисциплине, достаточную для дальнейшего изучения;

Оценка «хорошо (5)» дается студенту, продемонстрировавшему знание материала основной образовательной программы в объеме, необходимом для дальнейшего обучения и будущей работы по профессии, который, не проявляя достаточной активности на уроках, тем не менее самостоятельно выполнял овладел основными задачами, предусмотренными программой, освоил основную литературу, рекомендованную программой, допустил ошибки в их выполнении и ответе во время тестирования, но имеет необходимые знания для исправления этих ошибок самостоятельно;

Оценка «удовлетворительно (4)» дается студенту, обнаружившему знание материала основной образовательной программы в объеме, необходимом для дальнейшего обучения и будущей работы по профессии, который, не проявляя достаточной активности на уроках, тем не менее самостоятельно выполнял основные задачи, предусмотренные программой, изучил основную литературу, но допустил ошибки в их выполнении и в своем ответе во время теста, но имеет необходимые знания для исправления этих ошибок под руководством преподавателя;

Оценка «удовлетворительно (3)» выставляется обучающемуся, проявившему знание материала основной образовательной программы в объеме, необходимом для дальнейшего обучения и будущей работы по профессии, не проявившего активности на занятиях, самостоятельно выполнившего основные задания, предусмотренные законодательством. программа, но допускающая ошибки в их выполнении и в ответе во время теста, но обладающая необходимыми знаниями для устранения под руководством преподавателя наиболее существенных ошибок;

Оценка «неудовлетворительно (2)» дается студенту, который показал пробелы в знаниях или недостаток знаний по значительной части материала основной образовательной программы, не выполнил самостоятельно основные задачи, требуемые программой, допустил принципиальные ошибки в выполнении предусмотренных программой задач, не имеющего возможности продолжить учебу или начать профессиональную деятельность без дополнительной подготовки по данной дисциплине;

Оценка «неудовлетворительно (1)» ставится студенту при отсутствии ответа (отказ от ответа), либо когда представленный ответ совсем не соответствует сути вопросов, содержащихся в задании.

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

Во время экзамена студенту разрешается использовать программу дисциплины.