

**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)**

<b>по дисциплине:</b>	Глубокое обучение в прикладных задачах компьютерной лингвистики
<b>по направлению:</b>	Информатика и вычислительная техника
<b>профиль подготовки:</b>	Прикладная математика и информатика Физтех-школа Прикладной Математики и Информатики кафедра проблем передачи информации и анализа данных
<b>курс:</b>	1
<b>квалификация:</b>	магистр

Семестр, формы промежуточной аттестации: 2 (весенний) - Дифференцированный зачет

Аудиторных часов: 30 всего, в том числе:

лекции: 15 час.

семинары: 15 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 15 час.

Всего часов: 45, всего зач. ед.: 1

Программу составил: А.Н. Соболевский, д-р физ.-мат. наук, заведующий кафедрой

Программа обсуждена на заседании кафедры проблем передачи информации и анализа данных 31.05.2023

## Аннотация

Курс ставит целью познакомить студента с задачами компьютерной лингвистики и современными методами их решения, основанными на машинном обучении и, прежде всего, нейронных сетях. Предполагается обозреть основные задачи компьютерной лингвистики (текстовая классификация, морфологический и синтаксический анализ, распознавание именованных сущностей, определение семантической близости и т.д.), научить студента решать их с помощью стандартных средств (векторные представления слов, линейные модели, свёрточные и рекурсивные нейронные сети), а также пользоваться существующими программными средствами. К концу курса студент должен быть способен самостоятельно разбираться в современных работах по компьютерной лингвистике и реализовать изложенные в них методы.

## 1. Цели и задачи

### Цель дисциплины

- изучение современной компьютерной лингвистики, используемых в ней математических методов, обучение программированию компьютерно-лингвистических задач, а также подготовка слушателей к дальнейшей самостоятельной работе в области компьютерной лингвистики.

### Задачи дисциплины

- изучение современной компьютерной лингвистики,
- программирование компьютерно-лингвистических задач,
- подготовка к дальнейшей самостоятельной работе в области компьютерной лингвистики.

## 2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-2 Имеет представление об актуальных проблемах науки и техники в области информатики и вычислительной техники, 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 информатики и вычислительной техники 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
ПК-3 Владеет навыками участия в научных дискуссиях, выступления с сообщениями и докладами устного, письменного и виртуального (размещение в информационных сетях) характера, представления материалов собственных исследований	ПК-3.1 Знает основы ведения научной дискуссии и формы устного научного высказывания
	ПК-3.2 Умеет вести корректную дискуссию в области информационных технологий задавать вопросы и отвечать на поставленные вопросы по теме научной работы
	ПК-3.3 Имеет практический опыт участия в научных студенческих конференциях, очных, виртуальных, заочных обсуждениях научных проблем в области информационных технологий
ПК-1 Готов к включению в профессиональное сообщество; способен проводить под научным руководством локальные исследования на основе существующих методов в конкретной области профессиональной деятельности	ПК-1.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации; владеет навыками подготовки научных обзоров, публикаций, рефератов и библиографий по тематике проводимых исследований на русском и английском языке
	ПК-1.2 Умеет решать научные задачи с пониманием существующих подходов к верификации моделей программного обеспечения в связи с поставленной целью и в соответствии с выбранной методикой

### 3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- место компьютерной лингвистики среди задач искусственного интеллекта и её практические приложения;
- классификацию основных задач современной компьютерной лингвистики и их постановку при практической реализации;
- объекты теории формальных языков, используемые при решении задач компьютерной лингвистики (конечные автоматы, контекстно-свободные грамматики);
- основные типы нейронных сетей и то, к каким задачам лингвистики они применимы;
- математические основы автоматического обучения нейронных сетей.

уметь:

- сводить практическую задачу к одной или нескольким стандартным задачам компьютерной лингвистикой;
- самостоятельно подбирать алгоритм, наиболее подходящий для решения данной задачи;
- подбирать данные, необходимые для решения поставленной задачи;
- реализовать выбранный алгоритм на языке Python с использованием необходимых библиотек;
- оценивать качество реализации алгоритма, подбирать его оптимальные параметры.

владеть:

- основными библиотеками для машинного обучения и обработки естественного языка;
- навыками решения практических задач компьютерной лингвистики.

### 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.	1	1		1
2	Уровни обработки и представления текста.	2	2		1
3	Необходимые сведения из теории формальных языков.	2	2		1
4	Векторные представления слов.	1	1		1
5	Простейшие способы получения векторного представления текста.	1	1		1
6	Автоматический морфологический анализ.	2	2		1
7	Контекстные методы морфологического анализа.	1	1		1
8	Методы автоматического синтаксического анализа.	2	2		4
9	Задача распознавания именованных сущностей.	2	2		2
10	Морфологическая и синтаксическая разметка.	1	1		2
Итого часов		15	15		15

Подготовка к экзамену	0 час.
Общая трудоёмкость	45 час., 1 зач.ед.

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

Семестр: 2 (Весенний)

##### 1. Введение.

Задачи автоматической обработки текста и вычислительной лингвистики, их практические приложения.

##### 2. Уровни обработки и представления текста.

Токенизация, лемматизация, морфологический и синтаксический анализ.

##### 3. Необходимые сведения из теории формальных языков.

Регулярные выражения, конечные автоматы, контекстно-свободные грамматики.

##### 4. Векторные представления слов.

Векторные представления слов: word2vec, GloVe, FastText. Их применение в задачах вычислительной семантики: определение семантической близости.

##### 5. Простейшие способы получения векторного представления текста.

Методы усреднения и взвешивания слов: tf-idf и др. Простейшие задачи текстовой классификации и ранжирования. Применение дополнительной лингвистической информации для векторного представления текста.

##### 6. Автоматический морфологический анализ.

Лемматизация, определение морфологической метки. Бесконтекстные методы.

##### 7. Контекстные методы морфологического анализа.

Скрытые марковские модели, условные случайные поля.

##### 8. Методы автоматического синтаксического анализа.

Контекстно-свободные грамматики (грамматики составляющих), алгоритм Кока-Янгера-Касами.

Проект Universal Dependencies, принципы морфологической и синтаксической разметки. Грамматики зависимостей.

Алгоритмы Чу-Лю-Эдмондса и Нивре.

##### 9. Задача распознавания именованных сущностей.

Условные случайные поля для решения данной задачи.

##### 10. Морфологическая и синтаксическая разметка.

Применение морфологической и синтаксической разметки для извлечения информации из текста.

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

Стандартная учебная аудитория с мультимедийным проектором и экраном.

**6. Перечень рекомендуемой литературы**

Основная литература

1. Введение в искусственный интеллект [Текст] : конспект лекций : [для студентов, аспирантов вузов] / Д.В.Смолин .— М. : Физматлит, 2004 .— 208 с.

Дополнительная литература

1. Прикладная и компьютерная лингвистика [Текст], коллективная монография/под ред. И. С. Николаева, О. В. Митрениной, Т. М. Ландо, -М., ЛЕНАНД, 2017

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

Не используются

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

Не предусмотрено.

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

1. Рекомендуется своевременно выполнять практические задания для наилучшего усвоения материала и достижения более высокой оценки.
2. Для подготовки к итоговой аттестации по предмету лучше всего пользоваться материалами лекций.

**ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)**

<b>по направлению:</b>	Информатика и вычислительная техника
<b>профиль подготовки:</b>	Прикладная математика и информатика Физтех-школа Прикладной Математики и Информатики кафедра проблем передачи информации и анализа данных
<b>курс:</b>	<u>1</u>
<b>квалификация:</b>	магистр

Семестр, формы промежуточной аттестации: 2 (весенний) - Дифференцированный зачет

**Разработчик:** А.Н. Соболевский, д-р физ.-мат. наук, заведующий кафедрой

## 1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-2 Имеет представление об актуальных проблемах науки и техники в области информатики и вычислительной техники, 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 информатики и вычислительной техники 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
ПК-3 Владеет навыками участия в научных дискуссиях, выступления с сообщениями и докладами устного, письменного и виртуального (размещение в информационных сетях) характера, представления материалов собственных исследований	ПК-3.1 Знает основы ведения научной дискуссии и формы устного научного высказывания
	ПК-3.2 Умеет вести корректную дискуссию в области информационных технологий задавать вопросы и отвечать на поставленные вопросы по теме научной работы
	ПК-3.3 Имеет практический опыт участия в научных студенческих конференциях, очных, виртуальных, заочных обсуждениях научных проблем в области информационных технологий
ПК-1 Готов к включению в профессиональное сообщество; способен проводить под научным руководством локальные исследования на основе существующих методов в конкретной области профессиональной деятельности	ПК-1.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации; владеет навыками подготовки научных обзоров, публикаций, рефератов и библиографий по тематике проводимых исследований на русском и английском языке
	ПК-1.2 Умеет решать научные задачи с пониманием существующих подходов к верификации моделей программного обеспечения в связи с поставленной целью и в соответствии с выбранной методикой
	ПК-1.3 Имеет практический опыт выступлений и научной аргументации при анализе объекта научной и профессиональной деятельности

## 2. Показатели оценивания компетенций

В результате изучения дисциплины «Глубокое обучение в прикладных задачах компьютерной лингвистики» обучающийся должен:

### знать:

- место компьютерной лингвистики среди задач искусственного интеллекта и её практические приложения;
- классификацию основных задач современной компьютерной лингвистики и их постановку при практической реализации;
- объекты теории формальных языков, использующиеся при решении задач компьютерной лингвистики (конечные автоматы, контекстно-свободные грамматики);
- основные типы нейронных сетей и то, к каким задачам лингвистики они применимы;
- математические основы автоматического обучения нейронных сетей.

### уметь:

- сводить практическую задачу к одной или нескольким стандартным задачам компьютерной лингвистики;
- самостоятельно подбирать алгоритм, наиболее подходящий для решения данной задачи;
- подбирать данные, необходимые для решения поставленной задачи;
- реализовать выбранный алгоритм на языке Python с использованием необходимых библиотек;
- оценивать качество реализации алгоритма, подбирать его оптимальные параметры.

**владеть:**

- основными библиотеками для машинного обучения и обработки естественного языка;
- навыками решения практических задач компьютерной лингвистики.

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

С целью контроля освоения обучающимися учебного материала проводится устный опрос в начале занятия по теме прошлого занятия.

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

1. Уровни обработки и представления текста. Токенизация, лемматизация, морфологический и синтаксический анализ.
2. Регулярные выражения. Определение, основные свойства.
3. Конечные автоматы. Теорема Клини.
4. Векторные представления слов: word2vec, GloVe, FastText.
5. Применение векторных представлений в задаче семантической близости.
6. Простейшие способы получения векторного представления текста. Методы усреднения и взвешивания слов: tf-idf и др.
7. Автоматический морфологический анализ: лемматизация, определение морфологической метки. Бесконтекстные методы.
8. Контекстные методы морфологического анализа: скрытые марковские модели.
9. Контекстные методы морфологического анализа: условные случайные поля.
10. Методы автоматического синтаксического анализа. Контекстно-свободные грамматики, алгоритм Кока-Янгера-Касами.
11. Методы автоматического синтаксического анализа. Проект Universal Dependencies, принципы морфологической и синтаксической разметки. Грамматики зависимостей.
12. Теорема о существовании состоятельного решения уравнения правдоподобия. Состоятельность оценки максимального правдоподобия. Теорема об асимптотической нормальности решения уравнения правдоподобия.
13. Методы автоматического синтаксического анализа. Алгоритмы Чу-Лю-Эдмондса и Нивре.
14. Задача распознавания именованных сущностей. Условные случайные поля для решения данной задачи.
15. Применение морфологической и синтаксической разметки для извлечения информации из текста.

#### **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**

Во время проведения дифференцированного зачета обучающиеся могут пользоваться программой дисциплины. Дифференцированный зачет проводится в устной форме в виде опроса студентов по вопросам.