

**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**

	<b>Рабочая программа дисциплины (модуля)</b>
<b>по дисциплине:</b>	Математические методы анализа текстов
<b>по направлению:</b>	Информатика и вычислительная техника
<b>профиль подготовки:</b>	Прикладная математика и информатика Физтех-школа Прикладной Математики и Информатики кафедра интеллектуальных систем
<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: 60 час.

Всего часов: 90, всего зач. ед.: 2

Количество контрольных работ, заданий: 2

Программу составил: К.В. Воронцов, д-р физ.-мат. наук, заведующий кафедрой

Программа обсуждена на заседании кафедры интеллектуальных систем 20.01.2025

## Аннотация

В курсе рассматриваются основные задачи и математические методы обработки естественного языка. От студентов требуются знания курсов линейной алгебры, математического анализа, теории вероятностей, математической статистики, методов оптимизации, машинного обучения и нейронных сетей, языка программирования Python.

### 1. Цели и задачи

#### Цель дисциплины

Курс посвящен методам анализа текстов на основе статистики и машинного обучения.

#### Задачи дисциплины

- приобретение теоретических знаний в области методов анализа текстов;
- изучению классических и современных методов решения задач анализа текстов.

### 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
ОПК-4 Способен успешно реализовывать решение поставленной задачи, провести анализ результата и представить выводы, применяя знания и навыки в области математики, естественных наук и информационно-коммуникационных технологий	ОПК-4.1 Способен применять знания и навыки по использованию информационно-коммуникационных технологий для поиска и изучения научной литературы, применения прикладных программных продуктов
	ОПК-4.2 Способен применять знание информационно-коммуникационных технологий для решения поставленной задачи, формулирования выводов и оценки полученных результатов
	ОПК-4.3 Способен аргументировано выбирать способ проведения научного исследования
	ОПК-4.4 Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями

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

В результате освоения дисциплины обучающиеся должны

знать:

- фундаментальные понятия и теории методов анализа текстов;
- основные области применения этих методов.

уметь:

- подобрать подходящий метода для своей задачи, наиболее полно учитывающий её особенности.

владеть:

- навыками самостоятельной работы при решении типовых задач;
- культурой постановки и моделирования практически значимых задач;
- практикой исследования и решения теоретических и прикладных задач;
- навыками теоретического анализа реальных задач, решаемых с помощью методов анализа текстов.

#### 4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

##### 4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Предварительная обработка текста	1	1		7
2	Модели для работы с последовательностями	2	2		7
3	Синтаксический анализ	2	2		7
4	Классификация текстов	2	2		7
5	Вероятностные модели	2	2		8
6	Глубокие нейронные сети в анализе текстов	2	2		8
7	Онтологии, тезаурусы	2	2		8
8	Определение тональности текстов	2	2		8
Итого часов		15	15		60
Подготовка к экзамену		0 час.			
Общая трудоёмкость		90 час., 2 зач.ед.			

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

Семестр: 2 (Весенний)

###### 1. Предварительная обработка текста

Токенизация, лемматизация, выделение коллокаций, регулярные выражения.

###### 2. Модели для работы с последовательностями

Скрытая марковская модель, модели максимальной энтропии и условные случайные поля. Применение в задачах определения частей речи, выделения именованных сущностей, снятия омонимии.

###### 3. Синтаксический анализ

Методы синтаксического анализа. Область применения. Типы алгоритмов для синтаксического анализа. Восстановление после ошибок. Средства разработки анализаторов.

###### 4. Классификация текстов

Постановка задачи классификации текстов. Решение задачи классификации (предобработка и индексация, уменьшение размерности пространства признаков, построение и обучение классификатора с помощью методов машинного обучения, оценка качества классификации).

#### 5. Вероятностные модели

Модель языка, N-граммы, сглаживание, концепция шумного канала.

Применение в задачах исправления опечаток и машинного перевода.

#### 6. Глубокие нейронные сети в анализе текстов

Глубокие нейронные сети в анализе текстов.

Тематические модели, дистрибутивная семантика, векторные представления слов.

#### 7. Онтологии, тезаурусы

Онтологии, тезаурусы, выделение семантических связей. Работа с википедией.

#### 8. Определение тональности текстов

Определение тональности текстов. Методы классификации тональности. Машинное обучение с учителем. Машинное обучение без учителя. Метод, основанный на теоретико-графовых моделях.

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

Необходимое оборудование для лекций: компьютер и мультимедийное оборудование (проектор, звуковая система).

### **6. Перечень рекомендуемой литературы**

Основная литература

1. Speech and Language Processing. Dan Jurafsky and James H. Martin. 2-nd edition. 2009.

Дополнительная литература

1. Natural Language Processing with Python. Stewen Bird et. al. 2-nd edition. 2016.
2. Juravsky, Manning - Video lectures on natural language processing.

### **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 требуется

### **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, алгоритмы.

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 конспектам лекций, 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.

В рамках курса предполагается четыре практических задания, несколько домашних заданий и дифференцируемый зачет. Каждое задание и зачет оцениваются по пятибалльной шкале.

**ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)**

**по направлению:** Информатика и вычислительная техника  
**профиль подготовки:** Прикладная математика и информатика  
Физтех-школа Прикладной Математики и Информатики  
кафедра интеллектуальных систем  
**курс:** 1  
**квалификация:** магистр

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

**Разработчик:** К.В. Воронцов, 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
ОПК-4 Способен успешно реализовывать решение поставленной задачи, провести анализ результата и представить выводы, применяя знания и навыки в области математики, естественных наук и информационно-коммуникационных технологий	ОПК-4.1 Способен применять знания и навыки по использованию информационно-коммуникационных технологий для поиска и изучения научной литературы, применения прикладных программных продуктов
	ОПК-4.2 Способен применять знание информационно-коммуникационных технологий для решения поставленной задачи, формулирования выводов и оценки полученных результатов
	ОПК-4.3 Способен аргументировано выбирать способ проведения научного исследования
	ОПК-4.4 Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями

## 2. Показатели оценивания компетенций

В результате изучения дисциплины «Математические методы анализа текстов» обучающийся должен:

### знать:

- фундаментальные понятия и теории методов анализа текстов;
- основные области применения этих методов.

### уметь:

- подобрать подходящий метод для своей задачи, наиболее полно учитывающий её особенности.

### владеть:

- навыками самостоятельной работы при решении типовых задач;
- культурой постановки и моделирования практически значимых задач;
- практикой исследования и решения теоретических и прикладных задач;
- навыками теоретического анализа реальных задач, решаемых с помощью методов анализа текстов.

## 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 прошлой лекции или в конце занятия по пройденной теме.

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

1. Предварительная обработка текста.
2. Токенизация, лемматизация, выделение коллокаций, регулярные выражения.
3. Модели для работы с последовательностями.

4. Скрытая марковская модель, модели максимальной энтропии и условные случайные поля.
5. Применение в задачах определения частей речи, выделения именованных сущностей, снятия омонимии.
6. Синтаксический анализ.
7. Классификация текстов.
8. Вероятностные модели.
9. Модель языка, N-граммы, сглаживание, концепция шумного канала.
10. Применение в задачах исправления опечаток и машинного перевода.
11. Глубокие нейронные сети в анализе текстов.
12. Тематические модели, дистрибутивная семантика, векторные представления слов.
13. Онтологии, тезаурусы, выделение семантических связей. Работа с википедией.
14. Определение тональности текстов.

Примерный перечень билетов:

Билет №1

1. Токенизация, лемматизация, выделение коллокаций, регулярные выражения.
2. Применение в задачах исправления опечаток и машинного перевода.

Билет №2

1. Синтаксический анализ.
2. Онтологии, тезаурусы, выделение семантических связей. Работа с википедией.

#### Критерии оценивания

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

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

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

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

Оценка хорошо 6 баллов - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности.

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

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

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

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



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

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

Во время проведения дифференцированного зачета при подготовке билета можно пользоваться любыми материалами. При непосредственном ответе ничем пользоваться нельзя. Просьба обратить внимание на теоретический минимум. Незнание ответов на вопросы из теоретического минимума автоматически влечёт неудовлетворительную оценку за дифференцированный зачет.