

**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>	Вероятностные тематические модели
<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>	2
<b>квалификация:</b>	магистр

Семестр, формы промежуточной аттестации: 3 (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: К.В. Воронцов, д-р физ.-мат. наук

Программа обсуждена на заседании кафедры интеллектуальных систем 03.04.2024

## Аннотация

В курсе изучается вероятностное тематическое моделирование (topic modeling) коллекций текстовых документов. Тематическая модель определяет, какие темы содержатся в большой текстовой коллекции, и к каким темам относится каждый документ. Тематические модели позволяют искать тексты по смыслу, а не по ключевым словам, и создавать информационно-поисковые системы нового поколения, основанные на парадигме семантического разведочного поиска (exploratory search). Рассматриваются тематические модели для классификации, категоризации, сегментации, суммаризации текстов естественного языка, а также для рекомендательных систем, анализа банковских транзакционных данных, анализа биомедицинских сигналов. В спецкурсе развивается многокритериальный подход к построению моделей с заданными свойствами — аддитивная регуляризация тематических моделей (ARTM). Он основан на регуляризации некорректно поставленных задач стохастического матричного разложения. Особое внимание уделяется методам лингвистической регуляризации для моделирования связности текста. Предполагается проведение студентами численных экспериментов на модельных и реальных данных с помощью библиотеки тематического моделирования BigARTM.

### 1. Цели и задачи

#### Цель дисциплины

- изучение вероятностного тематического моделирования (topic modeling) коллекций текстовых документов.

#### Задачи дисциплины

- приобретение студентами навыков по методам анализа текстов и построения тематических моделей.

### 2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ПК-2 Понимает и способен применить в научно-исследовательской и прикладной деятельности основные законы естествознания, современный математический аппарат и алгоритмы, современные информационно-коммуникационные технологии	ПК-2.1 Знает основы научно-исследовательской деятельности в области информационных технологий, владеет знанием основ философии и методологии науки; знанием методов научных исследований и навыками их проведения
	ПК-2.2 Умеет применять полученные знания в области фундаментальных научных основ теории информации и решать стандартные задачи в собственной научно-исследовательской деятельности
	ПК-2.3 Имеет практический опыт научно-исследовательской деятельности в области информационно-коммуникационных технологий
	ПК-2.4 Владеет методами и алгоритмами решения задач цифровой обработки сигналов, использования сети Интернет, аннотирования, реферирования, библиографического поиска, опыт работы с научными источниками
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

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

В результате освоения дисциплины обучающиеся должны

знать:

- фундаментальные понятия, алгоритмы построения вероятностных тематических моделей;
- современные методы построения вероятностных тематических моделей.

уметь:

- пользоваться полученными знаниями для решения фундаментальных и прикладных задач;
- применять современные математические методы интеллектуального анализа данных;
- эффективно использовать информационные технологии и компьютерную технику для достижения необходимых теоретических и прикладных результатов.

владеть:

- культурой постановки и моделирования прикладных задач;
- практикой исследования и решения теоретических и прикладных задач;
- навыками теоретического анализа реальных задач интеллектуального анализа данных;
- навыками самостоятельной работы с литературой и в Интернете.

### 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	Задача построения тематической модели	3	3		6
2	Оценивание качества тематических моделей	9	9		23
3	Вероятностные тематические модели	9	9		23
4	Моделирование локального контекста	9	9		23
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: 3 (Осенний)

## 1. Задача построения тематической модели

Задача тематического моделирования.

Онлайновый ЕМ-алгоритм и регуляризаторы.

Разведочный информационный поиск.

Типичные приложения:

- анализ коллекций научных статей;
- анализ новостных потоков;
- рубрикация коллекций изображений, видео, музыки;
- аннотация генома и другие задачи биоинформатики;
- коллаборативная фильтрация.

## 2. Оценивание качества тематических моделей

Оценивание качества тематических моделей.

BigARTM и базовые инструменты.

Теория ЕМ-алгоритма.

## 3. Вероятностные тематические модели

Байесовское обучение модели LDA.

Тематические модели сочетаемости слов.

Анализ зависимостей.

Мультимодальные тематические модели.

## 4. Моделирование локального контекста

Моделирование локального контекста.

Суммаризация и визуализация.

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

Необходимое оборудование для лекций и практических занятий: компьютер и мультимедийное оборудование (проектор, маркерная доска, связь с Интернетом).

## **6. Перечень рекомендуемой литературы**

Основная литература

1. Воронцов К. В. Обзор вероятностных тематических моделей. 2018.

Дополнительная литература

1. Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman. Indexing by Latent Semantic Analysis // JASIS (41) 1990 pp. 391-407.
2. Thomas Hofmann. Probabilistic latent semantic analysis // Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval. 1999.
3. David M. Blei, Andrew Ng, Michael Jordan. Latent Dirichlet allocation // Journal of Machine Learning Research (3) 2003 pp. 993-1022.
4. T. L. Griffiths, M. Steyvers. Finding scientific topics // Proceedings of the National Academy of Sciences, Vol. 101, Nr. Suppl. 1 (April 2004) , p. 5228-5235. Скачать с CiteSeer
5. Mark Steyvers, Tom Griffiths. Probabilistic Topic Models // In Handbook of Latent Semantic Analysis. 2007.
6. Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad. Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of Computer Science in China, Vol.4, No.2, 2010, p. 280-301. Перевод на русский язык (PDF, 1 МБ).
7. Khoat Than, Tu bao Ho. Fully sparse topic models // Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD). 2012. Реферат статьи на русском языке (PDF, 1 МБ).

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

<http://www.machinelearning.ru/wiki/index.php>  
<http://www.machinelearning.ru/wiki/index.php?title=BigARTM>

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

Программное обеспечение и информационные технологии не требуются.

## **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:** Информатика и вычислительная техника  
**профиль подготовки:** Прикладная математика и информатика  
Физтех-школа Прикладной Математики и Информатики  
кафедра интеллектуальных систем  
**курс:** 2  
**квалификация:** магистр

Семестр, формы промежуточной аттестации: 3 (осенний) - Дифференцированный зачет

**Разработчик:** К.В. Воронцов, д-р физ.-мат. наук

## 1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ПК-2 Понимает и способен применить в научно-исследовательской и прикладной деятельности основные законы естествознания, современный математический аппарат и алгоритмы, современные информационно-коммуникационные технологии	ПК-2.1 Знает основы научно-исследовательской деятельности в области информационных технологий, владеет знанием основ философии и методологии науки; знанием методов научных исследований и навыками их проведения
	ПК-2.2 Умеет применять полученные знания в области фундаментальных научных основ теории информации и решать стандартные задачи в собственной научно-исследовательской деятельности
	ПК-2.3 Имеет практический опыт научно-исследовательской деятельности в области информационно-коммуникационных технологий
	ПК-2.4 Владеет методами и алгоритмами решения задач цифровой обработки сигналов, использования сети Интернет, аннотирования, реферирования, библиографического поиска, опыт работы с научными источниками
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 «Вероятностные тематические модели» обучающийся должен:

### знать:

- фундаментальные понятия, алгоритмы построения вероятностных тематических моделей;
- современные методы построения вероятностных тематических моделей.

### уметь:

- пользоваться полученными знаниями для решения фундаментальных и прикладных задач;
- применять современные математические методы интеллектуального анализа данных;
- эффективно использовать информационные технологии и компьютерную технику для достижения необходимых теоретических и прикладных результатов.

### владеть:

- культурой постановки и моделирования прикладных задач;
- практикой исследования и решения теоретических и прикладных задач;
- навыками теоретического анализа реальных задач интеллектуального анализа данных;
- навыками самостоятельной работы с литературой и в Интернете.

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

Условием сдачи курса является выполнение индивидуальных практических заданий в течении семестра

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

1. Задача тематического моделирования.
2. Онлайновый ЕМ-алгоритм и регуляризаторы.
3. Разведочный информационный поиск.
4. Оценивание качества тематических моделей.
5. BigARTM и базовые инструменты.
6. Теория ЕМ-алгоритма.
7. Байесовское обучение модели LDA.
8. Тематические модели сочетаемости слов.
9. Анализ зависимостей.
10. Мультимодальные тематические модели.
11. Моделирование локального контекста.
12. Суммаризация и визуализация.

#### **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 обучающиеся могут пользоваться программой дисциплины, а также справочной литературой, вычислительной техникой, конспектами лекций.

Дифференцированный зачет может проводиться по итогам текущей успеваемости и сдачи заданий, или путем организации специального опроса, проводимого в устной форме.