МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В.Ломоносова»

УТВЕРЖДАЮ

Декан факультета ВМК МГУ
имени М.В. Ломоносова

академик _____ Е.И. Моисеев

«____»_____ 2015 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

«Вероятностное тематическое моделирование»

Уровень высшего образования – подготовка научно-педагогических кадров в аспирантуре

Направление подготовки – 01.06.01 «Математика и механика»

Направленность (профиль) – 01.01.09 «Дискретная математика и математическая кибернетика»

2015

1. НАИМЕНОВАНИЕ ДИСЦИПЛИНЫ

Вероятностное тематическое моделирование

2. УРОВЕНЬ ВЫСШЕГО ОБРАЗОВАНИЯ

Подготовка научно-педагогических кадров в аспирантуре

3. НАПРАВЛЕНИЕ ПОДГОТОВКИ, НАПРАВЛЕННОСТЬ (ПРОФИЛЬ) ПОДГОТОВКИ

Направление подготовки – 01.06.01 «Математика и механика». Направленность (профиль) – 01.01.09 «Дискретная математика и математическая кибернетика»

4. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОСНОВНОЙ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Дисциплина относится к специальным дисциплинам вариативной части образовательной программы

5. ПЕРЕЧЕНЬ ПЛАНИРУЕМЫХ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ

Дисциплина участвует в формировании следующих компетенций образовательной программы:

Формируемые компетенции	Планируемые результаты обучения
Способностью формулировать научные задачи в области обеспечения информационной безопасности, применять для их решения методологии теоретических и экспериментальных научных исследований, внедрять полученные результаты в практическую деятельность (ОПК-1);	З1(ОПК-1) ЗНАТЬ научные задачи в области обеспечения информационной безопасности У1(ОПК-1) УМЕТЬ: применять для их решения методологии теоретических и экспериментальных научных исследований, внедрять полученные результаты в практическую деятельность В1(ОПК-1) ВЛАДЕТЬ: Навыками внедрения полученных результатов в практическую деятельность
Владение современными методами построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также методами разработки и реализации алгоритмов их решения на основе фун-	З1 (ПК-1) ЗНАТЬ: современные методы разработки и реализации алгоритмов организации работы вычислительных комплексов и компьютерных сетей последнего

даментальных знаний в области математики и информатики (ПК-1)	поколения У1 (ПК-1) УМЕТЬ: применять современные методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современные методы разработки и реализации алгоритмов их решения В1 (ПК-1) ВЛАДЕТЬ: навыками оптимального выбора современных методов построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методов разработки и реализации алгоритмов их решения

Оценочные средства для промежуточной аттестации приведены в Приложении.

6. ОБЪЕМ ДИСЦИПЛИНЫ

Объем дисциплины составляет 3 зачетные единицы, всего 108 часов.

24 часа составляет контактная работа с преподавателем – 22 часа занятий лекционного типа, 0 часов занятий семинарского типа (семинары, научно-практические занятия, лабораторные работы и т.п.), 0 часов индивидуальных консультаций, 0 часов мероприятий текущего контроля успеваемости, 0 часов групповых консультаций, 2 часа мероприятий промежуточной аттестации.

84 часов составляет самостоятельная работа аспиранта.

7. ВХОДНЫЕ ТРЕБОВАНИЯ ДЛЯ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Учащиеся должны владеть знаниями по дискретной математике и основам кибернетики в объеме, соответствующем основным образовательным программам бакалавриата и магистратуры по укрупненным группам направлений и специальностей 01.00.00 «Математика и механика», 02.00.00 «Компьютерные и информационные науки».

8. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

В процессе обучения технические и программные средства не используются.

9. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

В курсе излагаются основные результаты и методы теории сложности вычислений. Основное внимание уделяется классам сложностей и отношениям между ними.

Наименование и краткое содержание разделов и тем дисциплины (модуля), **форма промежуточной аттестации по дисциплине (модулю)**	**Всего (часы)**	В том числе								
		Контактная работа (работа во взаимодействии с преподавателем), часы из них						**Самостоятельная работа обучающегося, часы** из них		
		Занятия лекционного типа	Занятия семинарского типа	Групповые консультации	Индивидуальные консультации	Учебные занятия, направленные на проведение текущего контроля успеваемости (коллоквиумы, практические контрольные занятия и др)*	**Всего**	Выполнение домашних заданий	Подготовка рефератов и т.п.	**Всего**
Тема 1. Тематическая модель определяет, какие темы содержатся в большой текстовой коллекции, и к каким темам относится каждый документ. Материалы для первого ознакомления:	6	2					2		4	4

Тематический анализ больших данных. Краткое популярное введение в BigARTM. Разведочный информационный поиск. Видеолекция на ПостНауке. Тематическое моделирование. FAQ на ПостНауке, совместно с Корпоративным университетом Сбербанка. Тематическое моделирование на пути к разведочному информационному поиск										
Тема 2. 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). Построение кластеров	6	2					2		4	4

как решение задач оптимизации для внутренних индексов. Индексы Ball-Hall, Banfeld-Raftery, C, Calinski-Harabasz, Dunn, Silhouette. Внешние индексы Folkes-Mallows, Jaccard, Rand, Hungarian. Сравнения индексов.										
Тема 3.Аддитивная регуляризация тематических моделей. Понятие некорректно поставленной задачи по Адамару. Регуляризация. Теорема о необходимом условии максимума регуляризованного правдоподобия для ARTM. Условия Каруша–Куна–Таккера. Классические тематические модели PLSA и LDA как частные случаи ARTM. Мультимодальные тематические модели. Библиотека BigARTM.	6	2					2		4	4

Тема 4. Разведочный информационный поиск. Концепция разведочного поиска. Особенности разведочного поиска. Разведочный поиск как рекомендательная система. Часто используемые регуляризаторы. Сглаживание, разреживание, декоррелирование. Модальности. Иерархии тем. Послойное построение иерархии. Псевдодокументы родительских тем. Эксперименты с тематическим поиском. Методика измерения качества поиска. Тематическая модель для документного поиска. Оптимизация гиперпараметров.	6	2					2		4	4
Тема 5. Измерение	6	2					2		4	4

качества тематических моделей. Правдоподобие и перплексия. Интерпретируемость и когерентность. Разреженность и различность. Эксперименты с регуляризацией. Проблема определения числа тем. Проблема несбалансированности тем. Комбинирование регуляризаторов.										
Тема 6. Предварительная обработка текстов Парсинг "сырых" данных. Токенизация, стемминг и лемматизация. Выделение энграмм. Законы Ципфа и Хипса. Фильтрация словаря коллекции. Удаление стоп-слов. Библиотека BigARTM Методологические рекоммендации по про-	6	2					2		4	4

ведению экспериментов. Установка BigARTM. Формат и импорт входных данных. Обучение простой модели (без регуляризации): создание, инициализация, настройка и оценивание модели. Инструмент визуализации тем										
Тема 7. Классические модели PLSA, LDA. Модель PLSA. Модель LDA. Распределение Дирихле и его свойства. Максимизация апостериорной вероятности для модели LDA. Общий EM-алгоритм. EM-алгоритм для максимизации неполного правдоподобия. Регуляризованный EM-алгоритм. Сходимость в слабом смысле. Альтернативный вывод формул ARTM. Эксперименты с моделями PLSA, LDA.	8	2					2		6	6

Проблема неустойчивости (на синтетических данных). Проблема неустойчивости (на реальных данных). Проблема переобучения и робастные модели.										
Тема 8. Вариационный байесовский вывод. **Основная теорема вариационного байесовского вывода.** Вариационный байесовский вывод для модели LDA. VB EM-алгоритм для модели LDA. Сэмплирование Гиббса. Основная теорема о сэмплировании Гиббса. Сэмплирование Гиббса для модели LDA. GS EM-алгоритм для модели LDA.	8	2					2		6	6
Тема 9. Кусочно-	8	2					2		6	6

линейный метод восстановления зависимостей Кусочно-линейная регрессия, методы ее построения. Метод наименьших квадратов и его применение.										
Тема 10. Мультиграммные модели. Модель BigramTM. Модель Topical N-grams (TNG). Мультимодальная мультиграммная модель. Автоматическое выделение терминов. Алгоритм TopMine для быстрого поиска частых фраз. Критерии выделения коллокаций. Network Topic Model) и WTM (Word Topic Model). Связь с моделью word2vec. Понятие когерентности (согласованности). Экспериментально установленная связь ко-	8	2					2		6	6

герентности и интерпретируемости. Регуляризаторы когерентности.										
Тема 11. Зависимости, корреляции, связи. Тематические модели классификации и регрессии. Модель коррелированных тем CTM (Correlated Topic Model). Регуляризаторы гиперссылок и цитирования. Выявление тематических влияний в научных публикациях. Время и пространство. Регуляризаторы времени. Обнаружение и отслеживание тем. Гео-пространственные модели.	8	2					2		6	6
12. Промежуточная аттестация – устный экзамен	32	2						30		

Итого	108	24	84

10. УЧЕБНО-МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ УЧАЩИХСЯ

Самостоятельная работа учащихся состоит в изучении лекционного материала, учебно-методической литературы, подготовки к промежуточной аттестации.
Условием сдачи курса является выполнение индивидуальных практических заданий.
Рекомендуемая структура отчёта об исследовании по индивидуальному заданию:
Постановка задачи: неформальное описание, ДНК (дано–найти–критерий), структура данных
Описание простого решения baseline
Описание основного решения и его вариантов
Описание набора данных и методики экспериментов
Результаты экспериментов по подбору гиперпараметров основного решения
Результаты экспериментов по сравнению основного решения с baseline
Примеры визуализации модели
Выводы: что работает, что не работает, инсайты

Литература для самостоятельной работы студентов в соответствии с тематическим планом .

11. РЕСУРСНОЕ ОБЕСПЕЧЕНИЕ

Основная литература

1. Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. 2015.
2. Hofmann T. Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. — New York, NY, USA: ACM, 1999. — Pp. 50–57.
3. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation // Journal of Machine Learning Research. — 2003. — Vol. 3. — Pp. 993–1022.
4. Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models // Proceedings of the International Conference on Uncertainty in Artificial Intelligence. — 2009.
5. Янина А. О., Воронцов К. В. Мультимодальные тематические модели для разведочного поиска в коллективном блоге // Машинное обучение и анализ данных. 2015. Т.2. №2. С.173-186.
6. Ссылки

7. Тематическое моделирование
8. Аддитивная регуляризация тематических моделей
9. Коллекции документов для тематического моделирования
10. BigARTM
11. Видеозапись лекции на ТМШ, 19 июня 2015
12. Воронцов К.В. Практическое задание по тематическому моделированию, 2014.

Материально-техническая база

Для преподавания дисциплины требуется класс, оборудованный проектором и экраном.

12. ЯЗЫК ПРЕПОДАВАНИЯ

Русский

13. РАЗРАБОТЧИК ПРОГРАММЫ, ПРЕПОДАВАТЕЛИ

д. ф.-м. н. Воронцов Константин Вячеславович

14. ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ, ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИ

РЕЗУЛЬТАТ ОБУЧЕНИЯ	**КРИТЕРИИ и ПОКАЗАТЕЛИ ОЦЕНИВАНИЯ РЕЗУЛЬТАТА ОБУЧЕНИЯ из соответствующих карт компетенций**					**ОЦЕНОЧНЫЕ СРЕДСТВА**
	1	2	3	4	5	
	Неудовле-творитель-но	**Неудовлетворитель-но**	**Удовлетворительно**	**Хорошо**	**Отлично**	
УМЕТЬ: самостоятельно осуществлять научно-исследовательскую деятельность в соответствующей профессиональной области с использованием современных методов исследования и информационно-коммуникационных технологий У1 (ОПК-1)	Отсутствие умений	Частично освоенное умение самостоятельно осуществлять научно-исследовательскую деятельность в соответствующей профессиональной области с использованием современных методов исследования и информационно-коммуникационных технологий	В целом успешное, но не систематическое умение самостоятельно осуществлять научно-исследовательскую деятельность в соответствующей профессиональной области с использованием современных методов исследования и информационно-коммуникационных технологий	В целом успешное, но содержащее отдельные пробелы умение самостоятельно осуществлять научно-исследовательскую деятельность в соответствующей профессиональной области с использованием современных методов исследования и информационно-коммуникационных технологий	Успешное и систематическое умение самостоятельно осуществлять научно-исследовательскую деятельность в соответствующей профессиональной области с использованием современных методов исследования и информационно-коммуникационных технологий	доклад на научном семинаре
ЗНАТЬ:	Отсутствие	Фрагментарные пред-	В целом сформирован-	Сформированные, но	Сформированные	Устный экзамен

современные методы исследования и информационно-коммуникационных технологий в соответствующей профессиональной области З1(ОПК-1)	знаний	ставления современных методах исследования и информационно-коммуникационных технологий в соответствующей профессиональной области	ные, но неполные знания о современных методах исследования и информационно-коммуникационных технологий в соответствующей профессиональной области	содержащие отдельные пробелы знания о современных методах исследования и информационно-коммуникационных технологий в соответствующей профессиональной области	систематические знания о современ ных методах исследования и информационно-коммуникационных технологий в соответствующей профессиональной области	
ЗНАТЬ: современные методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современные методы разработки и реализации алгоритмов их решения З1 (ПК-1)	Отсутствие знаний	Фрагментарные представления о современных методах построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методах разработки и реализации алгоритмов их решения	В целом сформированные, но неполные знания о современных методах построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методах разработки и реализации алгоритмов их решения	Сформированные, но содержащие отдельные пробелы знания о современных методах построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методах разработки и реализации алгоритмов их решения	Сформированные систематические знания о современных методах построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методах разработки и реализации алгоритмов их решения	Устный экзамен
УМЕТЬ: применять современные методы построения и анализа математических моделей, возникающих	Отсутствие умений	Фрагментарные умения применять современные методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современные методы	В целом успешное, но не систематическое умение применять современные методы построения и анализа математических моделей, возникающих при решении естественнонаучных	Успешное, но содержащее отдельные пробелы умение применять современные методы построения и анализа математических моделей, возникающих при	Сформированное умение применять современные методы построения и анализа математических моделей, возникающих при решении естественнонаучных	Устный экзамен

при решении естественнонаучных задач, а также современные методы разработки и реализации алгоритмов их решения У1 (ПК-1)		разработки и реализации алгоритмов их решения	задач, а также современные методы разработки и реализации алгоритмов их решения	решении естественнонаучных задач, а также современные методы разработки и реализации алгоритмов их решения	задач, а также современные методы разработки и реализации алгоритмов их решения	
ВЛАДЕТЬ: навыками оптимального выбора современных методов построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методов разработки и реализации алгоритмов их решения В1 (ПК-1)	Отсутствие навыков	Фрагментарное владение навыками оптимального выбора современных методов построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методов разработки и реализации алгоритмов их решения	В целом успешное, но не полное владение навыками оптимального выбора современных методов построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методов разработки и реализации алгоритмов их решения	Успешное, но содержащее отдельные пробелы владение навыками оптимального выбора современных методов построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методов разработки и реализации алгоритмов их решения	Сформированное владение навыками оптимального выбора современных методов построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методов разработки и реализации алгоритмов их решения	Устный экзамен, контрольные работы