Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования

Московский государственный университет имени М.В. Ломоносова

Механико-математический факультет

Кафедра теоретической информатики

УТВЕРЖДАЮ

Заведующий кафедрой

\_\_\_\_\_\_\_\_\_\_\_\_\_\_/А.В.Михалев /

«22» января 2019 г.

**РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ**

**Наименование дисциплины :**

специальный курс по выбору студента

**Аналитика больших данных: дополнительные главы**

**Уровень высшего образования:**

**магистратура**

**Направление подготовки (специальность):**

* + 1. **Математика и компьютерные науки**

**Направленность (профиль) ОПОП:**

**Математика**

Форма обучения:

очная

Рабочая программа рассмотрена и одобрена

на заседании кафедры Теоретической информатики

(протокол №2а, «22 января 2019 года)

Москва 2019

Рабочая программа дисциплины разработана в соответствии с самостоятельно установленным МГУ образовательным стандартом (ОС МГУ) для реализуемых основных профессиональных образовательных программ высшего образования по специальности «Математика и компьютерные науки» (программы магистратуры) в редакции приказа МГУ от 30 декабря 2016 г.

Год (годы) приема на обучение\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**1.** Место дисциплины в структуре ОПОП ВО. **Вариативная часть ОПОП ВО.**

**2.** Входные требования для освоения дисциплины, предварительные условия: **знание** **основ статистики, линейной алгебры и** **программирования**.

**3.** Результаты обучения по дисциплине, соотнесенные с требуемыми компетенциями выпускников*.*

|  |  |
| --- | --- |
| **Компетенции выпускников (коды)** | **Планируемые результаты обучения по дисциплине , соотнесенные с компетенциями** |
| *ПК-1, ПК-5* | ***Знать:*** *стандартные методы обработки больших данных в задачах разработки рекомендательных систем, анализа социальных сетей, снижения размерности данных.****Уметь*** *ориентироваться в современных методах и алгоритмах компьютерной математики, уметь решать стандартные задачи по изученным темам.****Владеть:*** *прикладными методами решения практических задач по изученным темам.* |
| *ОПК-1, ОПК-2, ОПК-4* | ***Уметь*** *решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий****\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_******Иметь опыт*** *использования на практике математических алгоритмов, в том числе с применением современных вычислительных систем**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* |
| *УК-1* | ***Уметь*** *формулировать научно обоснованные гипотезы, создавать теоретические модели явлений и процессов, применять методологию научного познания в профессиональной деятельности****\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_******Владеть*** *специальными разделами фундаментальной математики, методами анализа и решения задач**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* |

**4.** Формат обучения очный.

**5.** Объем дисциплины составляет **4 з.е**., в том числе **50**  академических часов, отведенных на контактную работу обучающихся с преподавателем, **30**  академических часов на самостоятельную работу обучающихся.

**6.** Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий

|  |  |  |
| --- | --- | --- |
| **Наименование и краткое содержание разделов и тем дисциплины ,****Форма промежуточной аттестации по дисциплине**  | **Всего****(часы**) | В том числе |
| **Контактная работа (работа во взаимодействии с преподавателем)****Виды контактной работы, часы** | **Самостоятельная работа обучающегося,** **часы**  |
| Занятия лекционного типа\* | Занятия семинарского типа\* | **Всего** |  |
| Тема 1. Анализ ссылок в Интернет. | 4 | 2 |  | 2 | 2 |
| Тема 2. Вычисление PageRank. | 6 | 4 |  | 4 | 2 |
| Тема 3. Модифицированные алгоритмы вычисления PageRank. | 4 | 2 |  | 2 | 2 |
| Тема 4. Концентраторы и авторитеты, спам. | 4 | 2 |  | 2 | 2 |
| Тема 5. Рекомендательные системы, вызов NetFlix. | 6 | 4 |  | 4 | 2 |
| Тема 6. Алгоритмы классификации пользователей и товаров. | 4 | 2 |  | 2 | 2 |
| Текущий контроль успеваемости | 4 |  | 2 | 2 | 2 |
| Тема 7. Снижение размерности пространств данных. | 4 | 2 |  | 2 | 2 |
| Тема 8. Линейные алгоритмы снижения размерности, метод главных компонент. | 6 | 4 |  | 4 | 2 |
| Тема 9. Сингулярное разложение, CUR-декомпозиция. | 6 | 4 |  | 4 | 2 |
| Тема 10. Методы графовой кластеризации в социальных сетях. | 6 | 4 |  | 4 | 2 |
| Тема 11. Концепция машинного обучения в анализе больших данных. | 4 | 2 |  | 2 | 2 |
| Тема 12. Основные модели машинного обучения: линейные модели, деревья решений, нейронные сети. | 4 | 2 |  | 2 | 2 |
| Консультации | 4 | 2 | 2 |
| Экзамен/зачет | 4 | 4 |  |
| Промежуточная аттестация – контрольная работа | 10 | 8 | 2 |
| **Итого** | *80* |  *50* | 30 |

7. Фонд оценочных средств (ФОС) для оценивания результатов обучения по дисциплине

7.1. Типовые контрольные задания или иные материалы для проведения текущего контроля успеваемости.

1. Насколько разреженной должна быть матрица (то есть, какова в ней должна быть доля единичных элементов) для того, чтобы разреженное представление было более экономичным?

2. Используя компактный метод представления (источник, степень, преемники) представьте матрицы перехода определенных графов (в задании).

3. Матрица предпочтений представляет рейтинги по шкале 1-5 для восьми товаров, от *a* до *h*, для трех пользователей *A*, *B* и *C*. Вычислить следующие величины по данным из этой матрицы.

(а) Преобразуйте матрицу предпочтений в булев вид (есть оценка или ее нет) и вычислите расстояние по Жаккару между каждой парой пользователей.

(б) Повторите п. (а) для расстояния по косинусу.

(с) Преобразуйте рейтинги 3, 4 и 5 в 1; а 1, 2 и пустое значение – в 0. Вычислите расстояние по Жаккару между каждой парой пользователей.

(d) Повторите п. (с) для расстояния по косинусу.

(е) Нормализуйте матрицу вычитанием из каждого непустого значения среднего значения для данного пользователя.

(f) Используя нормализованную матрицу из п. (d), вычислить расстояние по косинусу между каждой парой пользователей.

4. Рассмотрим вершину в дереве решений с 80 экземплярами типа A и 70 экземплярами типа B. Вычислить энтропию данной вершины.

5. Рассмотрим следующий журнал событий L. С помощью альфа-алгоритма нарисуйте сеть Петри, соответствующую журналу L.

6. Рассмотрим следующий журнал событий L. Постройте матрицу следа (Footprint of L) и найдите коэффициент подтверждения на основе матрицы следа (Footprint-based conformance).

7. Рассмотрим следующую модель процесса РN и след < a,b,c,h,i >. Найти коэффициент подтверждения на основе выравнивания (correct alignment-based conformance) для этого следа, предполагая, что стоимость переходов для всех активностей равна 1.

7.2. Типовые контрольные задания или иные материалы для проведения промежуточной аттестации.

PageRank.

Эффективное вычисление PageRank.

Эффективное представление матрицы перехода.

PageRank, чувствительный к теме

TrustRank.

Концентраторы и авторитеты.

Рекомендательные системы.

Матрица предпочтений.

Рекомендации на основе содержания.

Обнаружение особенностей документов.

Алгоритмы классификации.

Кластеризация пользователей и товаров.

Снижение размерности.

UV-разложение.

Оптимизация произвольного элемента.

Метод главных компонент.

Матрица расстояний.

Сингулярное разложение.

Запросы с использованием концептов.

Вычисление сингулярного разложения матрицы.

CUR-декомпозиция.

Методы графовой кластеризации в социальных сетях.

Дерево принятия решений, понятие энтропии.

Кластерный анализ в  интеллектуальном анализе процессов.

Оценка результатов для дерева решений.

Теорема CAP.

Основные модели машинного обучения: линейные модели, деревья решений, нейронные сети.

|  |
| --- |
| **ШКАЛА И КРИТЕРИИ ОЦЕНИВАНИЯ результатов обучения (РО) по дисциплине (модулю)**  |
| ОценкаРО исоответствующие виды оценочных средств  | 2 | 3 | 4 | 5 |
| **Знания***(виды оценочных средств: устные и письменные опросы и контрольные работы, тесты, и т.п. )* | Отсутствие знаний | Фрагментарные знания | Общие, но не структурированные знания | Сформированные систематические знания |
| **Умения***(виды оценочных средств: практические контрольные задания, написание и защита рефератов на заданную тему и т.п.)*  | Отсутствие умений | В целом успешное, но не систематическое умение | В целом успешное, но содержащее отдельные пробелы умение (допускает неточности непринципиального характера) | Успешное и систематическое умение |
| **Навыки (владения, опыт деятельности)***(виды оценочных средств: выполнение и защита курсовой работы, отчет по практике, отчет по НИР и т.п.)*  | Отсутствие навыков (владений, опыта) | Наличие отдельных навыков (наличие фрагментарного опыта) | В целом, сформированные навыки (владения), но используемые не в активной форме | Сформированные навыки (владения), применяемые при решении задач |

8. Ресурсное обеспечение:

* Перечень основной и дополнительной литературы,
* А. Основная литература
* 1. Ю. Лесковец, А. Раджарамаран, Дж. Ульман. Анализ больших наборов данных. ДМК, Москва, 2016.
* 2. Wil M. P., van der Aalst. Process Mining Discovery, Conformance and Enhancement of Business Processes. ISBN: 978-3-642-19344-6 (Print) 978-3-642-19345-3 (Online).
* 3. Hasso Plattner. A Course in In-Memory Data Management. The Inner Mechanics of In-Memory Databases. ISBN:  978-3-642-55269-4 (Print)  978-3-642-55270-0 (Online).
* MongoDB Data Modeling. Focus on data usage and better design schemas with the help of MongoDB. Wilson da Rocha França. 2015, Packt Publishing. ISBN 978-1-78217-534-6.
* Б. Дополнительная литература

# 1. [Jure Leskovec](http://www.cambridge.org/ru/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/mining-massive-datasets-2nd-edition?format=HB#bookPeople), [Anand Rajaraman](http://www.cambridge.org/ru/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/mining-massive-datasets-2nd-edition?format=HB#bookPeople), [Jeffrey D. Ullman](http://www.cambridge.org/ru/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/mining-massive-datasets-2nd-edition?format=HB#bookPeople). Mining of Massive Datasets. 2nd Edition. Stanford University, California, 2014.

* Перечень ресурсов информационно-телекоммуникационной сети «Интернет»
1. <http://www.cambridge.org/ru/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/mining-massive-datasets-2nd-edition#c3bEPha9lWJHRUhW.99>
2. <http://i.stanford.edu/~ullman/mmdsn.html>
* Описание материально-технического обеспечения.
* Компьютерный класс

9. Язык преподавания.

Русский

10. Преподаватели

Главацкий С.Т., Бурыкин И.Г., Айдагулов Р.Р.

11. Авторы программы

Главацкий С.Т., Бурыкин И.Г., Айдагулов Р.Р.