МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

«Дальневосточный федеральный университет»

(ДВФУ)

ШКОЛА ЕСТЕСТВЕННЫХ НАУК

«СОГЛАСОВАНО»
Руководитель ОП Разработка программно-информационных систем по направлению 09.04.04 Программная инженерия

_____ Артемьева И.Л.
(подпись) (Ф.И.О. рук. ОП)
«21» 07 2018 г.

«УТВЕРЖДАЮ»
Заведующая кафедрой прикладной математики, механики, управления и программного обеспечения

_____ Артемьева И.Л.
(подпись) (Ф.И.О. зав. каф.)
«21» 07 2018 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

Основы аналитики больших объемов данных

Направление подготовки – 09.04.04 Программная инженерия

Магистерская программа «Разработка программно-информационных систем»

Форма подготовки (очная)

курс 1 семестр 2
лекции 18 час.
практические занятия 18 час.
лабораторные работы 0 час.
в том числе с использованием МАО лек. 0 / пр. 0/ лаб. 0 час.
в том числе в электронной форме лек _____/пр._____/лаб._____ час.
всего часов аудиторной нагрузки – 36 час.
в том числе с использованием МАО – 18 час.
в том числе контролируемая самостоятельная работа __ час.
в том числе в электронной форме _____ час.
самостоятельная работа 72 час.
в том числе на подготовку к экзамену _____ час
курсовая работа / курсовой проект не предусмотрено
зачет 2 семестр
экзамен не предусмотрено

Рабочая программа составлена в соответствии с требованиями образовательного стандарта, самостоятельно устанавливаемого ДВФУ, утвержденного приказом ректора от 07.07.2015 № 12-13-1282

Рабочая программа обсуждена на заседании кафедры прикладной математики, механики, управления и программного обеспечения, протокол № 7.2 от 21.07.2018 г.

Заведующая кафедрой прикладной математики, механики, управления и программного обеспечения д.т.н., профессор Артемьева И.Л.

Составитель: доцент кафедры прикладной математики, механики, управления и программного обеспечения Смагин С.В., к.т.н.

Оборотная сторона титульного листа РПУД

I. Рабочая программа пересмотрена на заседании кафедры:

Протокол от «_____» _____ 20___ г. № _____

Заведующий кафедрой _____ _____
(подпись) (И.О. Фамилия)

II. Рабочая программа пересмотрена на заседании кафедры:

Протокол от «_____» _____ 20___ г. № _____

Заведующий кафедрой _____ _____
(подпись) (И.О. Фамилия)

ABSTRACT

Master's degree in 09.04.04 – Software engineering

Master's Program "Development of software and information systems**"**

Course title: Fundamentals of large data analytics

Variable) part of Block, 3 **credits**

Instructor: Smagin S.

At the beginning of the course a student should be able to: study independently, be self-organized; know about main concepts, principles, theories and facts related to computer science; carry out the search, the storage, the treatment and the analysis of information from various sources and databases represent this information in a required form with the help of information, computer and network technologies; formalize own subject domain taking into account the restrictions of used research methods; formalize the subject domain of a project and develop specifications for the components of software.

Learning outcomes: an ability to perceive mathematical, naturally-scientific, social and economic and professional knowledge, an ability to independently get, develop and use it for solving nonstandard problems including the problems from a new or unknown field and an interdisciplinary context; culture of thinking, an ability to form the logic of reasoning and statements based on an interpretation of data integrated from various spheres of science and technology, to make judgments using incomplete data; possession of methods and tools of receiving, storage, processing and broadcasting of information by means of modern computer technologies including global computer networks; possession of the existing methods and algorithms of solving the problems of data recognition and processing; possession of the existing methods and algorithms of solving the problems of digital signal processing.

Course description: Modern methods of intellectual data analysis, methods of formation and assessment analysis of its internal and external properties.

Main course literature:

1. Bolotova L.S. Sistemy iskusstvennogo intellekta: modeli i tekhnologii, osnovannye na znaniyah: uchebnik [Systems of artificial intelligence: models and technologies based on knowledge: a textbook] / FGBOU VPO RGUITP; FGAU GNII ITT "Informika". – M.: Finansy i statistika, 2012. – 664 p. (rus) – access: http://www.studentlibrary.ru/book/ISBN9785279035304.html
2. Zagoruyko N.G. Kognitivnyy analiz dannyh. [Cognitive analysis of data] – Novosibirsk: Geo, 2013. – 183 p. (rus)

3. Vagin V.N. Dostovernyy i pravdopodobnyy vyvod v intellektual'nyh sistemah: uchebnoe posobie. [A reliable and plausible conclusion in intellectual systems: a tutorial] – M.: Fizmatlit, 2008. – 704 p. (rus) – access: http://znanium.com/go.php?id=544735
4. Osipov G.S. Metody iskusstvennogo intellekta. [Methods of artificial intelligence] – M.: FIZMATLIT, 2011. – 296 p. (rus) – access: http://www.studentlibrary.ru/book/ISBN9785922113236.html
5. Nizametdinov SH.U., Rumyancev V.P. Analiz dannyh: uchebnoe posobie. [Data Analysis: A Training Manual] – M.: NIYAU “MIFI”, 2012. – 288 p. (rus) – access: http://znanium.com/bookread2.php?book=567083

Form of final knowledge control: Pass-fail exam.

Аннотация рабочей программы учебной дисциплины
«Основы аналитики больших объемов данных»

Рабочая программа дисциплины «Основы аналитики больших объемов данных» разработана для студентов 1 курса, обучающихся по направлению 09.04.04 Программная инженерия, магистерская программа «Разработка программно-информационных систем». Дисциплина является обязательной дисциплиной вариативной части учебного плана Б1.В.03.01.

Трудоемкость дисциплины 3 зачетных единицы (108 часов). В 3 семестре дисциплина содержит 18 часов лекций, 0 лабораторных занятий, 18 часов практических занятий, 72 часа самостоятельной работы студентов.

Дисциплина «Основы аналитики больших объемов данных» базируется на дисциплине бакалавриата «Теория вероятностей и математическая статистика». Знания, полученные при ее изучении, будут использованы в дисциплинах «Методы машинного обучения», «Интеллектуальный анализ данных» учебного плана. Дисциплина реализуется в 3 семестре (семестрах).

Цель дисциплины – изучение современных методов интеллектуального анализа данных, а также способов формирования и анализа оценок их внешних и внутренних свойств.

Задачи дисциплины:

1. Изучение алгоритмов обработки данных, применяемых для случая больших данных
2. Изучение особенностей этих алгоритмов и методов их применения.
3. Изучение методов сравнения алгоритмов и подготовки альтернативных решений.

Для успешного изучения дисциплины «Основы аналитики больших объемов данных» у обучающихся должны быть сформированы следующие предварительные компетенции:

- способность к самоорганизации и самообразованию;
- владение основными концепциями, принципами, теориями и фактами, связанными с информатикой;
- способность осуществлять поиск, хранение, обработку и анализ информации из различных источников и баз данных, представлять ее в требуемом формате с использованием информационных, компьютерных и сетевых технологий;
- способность к формализации в своей предметной области с учетом ограничений используемых методов исследования;

- способность формализовать предметную область программного проекта и разработать спецификации для компонентов программного продукта.

Планируемые результаты обучения по данной дисциплине (знания, умения, владения), соотнесенные с планируемыми результатами освоения образовательной программы, характеризуют этапы формирования следующих компетенций (общекультурные/ общепрофессиональные/ профессиональные компетенции (элементы компетенций)):

Код и формулировка компетенции	**Этапы формирования компетенции**	
ОК-4 умением быстро осваивать новые предметные области, выявлять противоречия, проблемы и вырабатывать альтернативные варианты их решения	Знает	Особенности существующих алгоритмов и технологий обработки данных.
	Умеет	Выявлять противоречия алгоритмов при их использовании для конкретных задач.
	Владеет	Методами адаптации алгоритмов для альтернативных решений.
ПК-2 знанием методов научных исследований и владением навыками их проведения	Знает	Методы поиска литературы по новым алгоритмам и технологиям обработки больших объемов данных.
	Умеет	Выделять в алгоритмах основное.
	Владеет	Методами сравнения алгоритмов.
ПК-4 владением существующими методами и алгоритмами решения задач распознавания и обработки данных	Знает	Основные алгоритмы решения задач распознавания и обработки данных.
	Умеет	Применять алгоритмы при анализе больших объемов данных.
	Владеет	Методами выбора подходящих алгоритмов для конкретных типов задач.

Для формирования вышеуказанных компетенций в рамках дисциплины «Основы аналитики больших объемов данных» применяются следующие методы активного/ интерактивного обучения: метод круглого стола, метод проектов.

I. СТРУКТУРА И СОДЕРЖАНИЕ ТЕОРЕТИЧЕСКОЙ ЧАСТИ КУРСА

Лекционный материал (18 часов)

Тема 1. Введение в большие данные (2 час.)

- Определение больших данных и причины их появления.
- Примеры возможностей для бизнеса.

Тема 2. Жизненный цикл аналитики данных (3 час.)

- Различие между Business Intelligence и Big Data.
- Понятие жизненного цикла аналитики данных.
- Роли, необходимые для успешного создания проекта по аналитике данных.
- Инструменты получения и обмена данными.

Тема 3. Высокопроизводительные вычисления (3 час.)

- Распределенные вычисления на нескольких серверах, вычислительная парадигма MapReduce.
- Проект Apache Hadoop и его экосистема.
- Apache Spark и его компоненты.
- Вычисления в реальном времени, Apache Storm, Flink.

Тема 4. Масштабирование и многоуровневое хранение данных (3 час.)

- Теорема CAP
- Парадигма NoSQL
- Классификация NoSQL баз данных

Тема 5. Визуализация данных и результатов анализа (3 час.)

- Техники визуализации данных.
- Язык R.

Тема 6. Сложные методы аналитики (2 час.)

- Классификация задач анализа: Text, Data, Web, Social Mining.
- Статистические методы анализа данных. Применение машинного обучения в аналитике.

Тема 7. Анализ текста (2 час.)

- Поисковые механизмы: Lucene, Solr, Elasticsearch.
- Алгоритм Work2Vec.

II. СТРУКТУРА И СОДЕРЖАНИЕ ПРАКТИЧЕСКОЙ ЧАСТИ КУРСА

Практические занятия (18 час.)

Практическое занятие 1. Выбор предметной области (4 час.)

Поставленная перед слушателями задача не привязана к какой-либо конкретной предметной области. Предполагается отойти от принципа выполнения заранее поставленных и четко сформулированных задач, чтобы предоставить исполнителю гибкость и возможность творческого подхода выполнения. Таким образом, исполнителю предоставляется возможность самостоятельного выбора интересующей его прикладной области, над которой в рамках курса будет проводиться работа.

Практическое занятие 2. Формирование набора данных (4 час.)

Во время выполнения проекта может потребоваться работать с информацией разного типа. Традиционно принято выделять четыре типа данных: структурированные данные, полуструктурированные данные, квазиструктурированные данные и не структурированные данные. Исполнитель самостоятельно выбирает тип данных, с которым в дальнейшем будет работать, но требуется принимать во внимание, что поскольку в курсе рассматриваются подходы и технологии обработки именно большого объема данных, то для выбранной прикладной области рекомендуется иметь для проведения анализа не менее 2 Гб структурированных или полуструктурированных данных, если не используются методы анализа неструктурированного контента. В случае, если используются методы анализа неструктурированного контента, такого как изображения, аудио- и видеозаписи, то рекомендуемый минимальный объем информации – 5Гб.

Практическое занятие 3. Архитектура проектируемой системы (4 час.)

Разрабатываемый проект подразумевает создание программного решения, позволяющего автоматически или полуавтоматически решать сформулированные задачи анализа. В основе решения может быть заложена относительно простая, но функциональная и расширяемая модульная схема. Стоит отметить, что программное решение должно обладать хорошей производительностью, гибким масштабированием, быть распределенным и гарантировать надежность передачи данных между узлами системы. Еще одной важной особенностью рассматриваемой архитектуры является возможность гибкой настройки проводимых в системе аналитик.

Практическое занятие 4. Хранение и обработка данных (3 час.)

Хранение сформированного набора данных или набора данных, который прошел предварительную очистку и готов поступить на обработку, предполагается осуществлять в базе данных.

Модуль обработки данных является одним из центральных модулей программной системы анализа. В нем заложена основная логика получения ответа на поставленную задачу анализа. Выбор используемых технологий и методов решения зависит именно от задачи. Для одного типа задач хорошо подходит использование методов машинного обучения и нейронных сетей, а для другого типа задач идеальным образом становится простое решение с помощью SQL-подобных запросов. Таким образом, выбор того, какую технологию использовать, ложится на слушателя.

Практическое занятие 5. Визуализация результатов (3 час.)

Результаты проведенных исследований над данными обычно представляются в виде набора графиков и диаграмм, наглядно изображающих полученные выводы. Возможно использование подходов изображения результатов в виде инфографики или облака тегов.

Для визуализации результатов аналитики разумно применение языка R. Он может быть использован для обработки «сырых» результатов анализа, объем которых не превышает 200…300 Мб, так как программы, созданные на нем (языке R), являются относительно медленными и не масштабируемыми, хотя на сегодняшний день и есть поддержка языка R в Spark для исполнения на кластере. Программы на языке R могут быть использованы для быстрого и наглядного изображения промежуточных результатов, на основе которых делается выбор дальнейшего направления движения при ответе на поставленный в задаче вопрос.

Лабораторные работы (0 час.)

Ш. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ ОБУЧАЮЩИХСЯ

Трудоемкость самостоятельной работы 72 часа. Учебно-методическое обеспечение самостоятельной работы обучающихся по дисциплине «Интеллектуальный анализ данных» представлено в Приложении 1 и включает в себя:

- план-график выполнения самостоятельной работы по дисциплине, в том числе примерные нормы времени на выполнение по каждому заданию;

- характеристика заданий для самостоятельной работы обучающихся и методические рекомендации по их выполнению;
- требования к представлению и оформлению результатов самостоятельной работы;
- критерии оценки выполнения самостоятельной работы.

IV. КОНТРОЛЬ ДОСТИЖЕНИЯ ЦЕЛЕЙ КУРСА

№ п/п	Контролируемые разделы/темы дисциплины	Коды и этапы формирования компетенций		Оценочные средства – наименование	
				текущий контроль	промежуточная аттестация
1	Темы 1-2	ОК-4 ПК-2 ПК-4	Знает	Собеседование УО-1, круглый стол УО-4	Зачет Вопросы 1-4
			Умеет	Практическое занятие 1 ПР-6	
2	Темы 3-4	ОК-4 ПК-2 ПК-4	Знает	Собеседование УО-1, круглый стол УО-4	Зачет Вопросы 5-7
			Умеет	Практическое занятие 2 ПР-6	
3	Темы 5-6	ОК-4 ПК-2 ПК-4	Знает	Собеседование УО-1, круглый стол УО-4	Зачет Вопросы 8-10
			Умеет	Практические занятия 3-5 ПР-6	

Типовые контрольные задания, методические материалы, определяющие процедуры оценивания знаний, умений и навыков и (или) опыта деятельности, а также критерии и показатели, необходимые для оценки знаний, умений, навыков и характеризующие этапы формирования компетенций в процессе освоения образовательной программы, представлены в Приложении 2.

V. СПИСОК УЧЕБНОЙ ЛИТЕРАТУРЫ И ИНФОРМАЦИОННО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Основная литература

1. Болотова Л.С. Системы искусственного интеллекта: модели и технологии, основанные на знаниях: учебник / ФГБОУ ВПО РГУИТП; ФГАУ ГНИИ ИТТ «Информика». – М.: Финансы и статистика, 2012. – 664 с.: ил. http://www.studentlibrary.ru/book/ISBN9785279035304.html
2. Много цифр. Анализ больших данных при помощи Excel / Форман Д.; Пер. с англ. Соколовой А. – М.: Альпина Пабл., 2016. – 461 с.: 84x108 1/16 ISBN 978-5-9614-5032-3 http://znanium.com/catalog/product/551044 http://www.studentlibrary.ru/book/ISBN9785961450323.html

3. Hadoop в действии / Чак Лэм – М. : ДМК Пресс, 2012. – 424 с. – ISBN 978-5-94074-785-7 http://www.studentlibrary.ru/book/ISBN9785940747857.html
4. Вагин В.Н. Достоверный и правдоподобный вывод в интеллектуальных системах: учебное пособие. – М.: Физматлит, 2008. – 704 с. http://znanium.com/go.php?id=544735
5. Низаметдинов Ш.У., Румянцев В.П. Анализ данных: учебное пособие. – М.: НИЯУ «МИФИ», 2012. – 288 с. ISBN 978-5-7262-1687-4 http://znanium.com/bookread2.php?book=567083

Дополнительная литература

1. Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим. Language Arts & Disciplines, 2013. – 599 с.
2. Силен Д., Мейсман А., Али М. Основы Data Science и Big Data. Python и наука о данных. СПб: Питер, 2017. – 336 стр.
3. Уайт Т. Hadoop: Подробное руководство. СПб: Питер, 2013. – 672 с.
4. Лэм Ч. Hadoop в действии. Москва: ДМК Пресс, 2012. – 426 с.
5. Фаулер М., Прамодкумар Дж. Садаладж. NoSQL: новая методология разработки нереляционных баз данных. – М.: «Вильямс», 2013. – 192 с.
6. Смородин В. В., Волкова Е. В., Алиев А. А. От хранения данных к управлению информацией. – СПб.: Изд-во Питер, 2010. – 528 с.
7. Яу Н. Искусство визуализации в бизнесе. Как представить сложную информацию простыми образами. – Wiley Publishing, Inc. 2013.
8. Храмов Д.А. Сбор данных в Интернете на языке R, 2016. – 282 с.
9. Шитиков В. К., Мастицкий С. Э. Классификация, регрессия, алгоритмы Data Mining с использованием R. – Электронная книга, адрес доступа: https://github.com/ranalytics/data-mining – 2017.

Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

1. http://www.inftech.webservis.ru/it/database/datamining/ar2.html Дюк В.А. Data Mining – интеллектуальный анализ данных.
2. http://kek.ksu.ru/EOS/dm.pdf Степанов Р.Г. Технология Data Mining: Интеллектуальный анализ данных / Казань, 2008.
3. http://machinelearning.ru/ MachineLearning.ru Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных.

Перечень информационных технологий и программного обеспечения

Для выполнения лабораторных работ требуется следующее программное обеспечение: Microsoft Excel, RGui, RStudio.

VI. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Дисциплина «Основы аналитики больших объемов данных» изучается в следующих организационных формах: лабораторная работа; самостоятельное изучение теоретического материала; индивидуальные и групповые консультации. Основной формой самостоятельной работы студента является изучение конспекта лекций, их дополнение рекомендованной литературой, выполнение проекта, а также активная работа на лабораторных занятиях.

К прослушиванию лекции следует готовиться, для этого необходимо знать программу курса и рекомендованную литературу. Тогда в процессе лекции легче отделить главное от второстепенного, легче сориентироваться: что записать, что самостоятельно проработать, что является трудным для понимания, а что легко усвоить. Контроль за выполнением самостоятельной работы студента производится в виде контроля каждого этапа работы, отраженного в документации и защиты проекта.

Студент должен планировать график самостоятельной работы по дисциплине и придерживаться его.

VII. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Лекции проводятся с использованием проектора и внутренней системы портала ДВФУ. Лабораторные занятия проходят в аудиториях, оборудованных компьютерами типа Lenovo C360G-i34164G500UDK с лицензионными программами Microsoft Office 2013 и аудио-визуальными средствами проектор Panasonic DLPProjectorPT-D2110XE, плазма LG FLATRON M4716CCBAM4716CJ. Для выполнения самостоятельной работы студенты в жилых корпусах ДВФУ обеспечены Wi-Fi.

Приложение 1

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

«Дальневосточный федеральный университет»

(ДВФУ)

ШКОЛА ЕСТЕСТВЕННЫХ НАУК

УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ ОБУЧАЮЩИХСЯ

по дисциплине «Основы аналитики больших объемов данных»

Направление подготовки – 09.04.04 Программная инженерия

Магистерская программа «Разработка программно-информационных систем»

Форма подготовки (очная)

Владивосток

2018

План-график выполнения самостоятельной работы по дисциплине

№ п/п	Формулировка задачи	Дата/сроки выполнения	Примерные нормы времени на выполнение	Форма контроля
1.	Формулировка задачи, выбор предметной области	7 неделя обучения	6 часа	Собеседование
2.	Разработка модели предметной области и исследование ее свойств	8-10 неделя обучения	14 часов	Собеседование
3.	Разработка метода интеллектуального анализа данных для разработанной модели, проведение тестирования	10-13 неделя обучения	14 часов	Проверка отчетов, собеседование
4.	Проведение компьютерного эксперимента на модельных данных	14-15 неделя обучения	14 часов	Проверка отчетов, собеседование
5.	Оценка внешних и внутренних свойств метода интеллектуального анализа данных, формирование заключения о качестве метода	16-17 неделя обучения	14 часов	Зачет
		всего	72 часа	

Рекомендации по самостоятельной работе студентов

Трудоемкость самостоятельной работы 72 часа. Самостоятельная работа обучающихся подразумевает обязательную подготовку к лабораторным занятиям (оформление отчетов), изучение основной и дополнительно литературы по дисциплине, подготовку к текущему контролю и промежуточной аттестации в конце семестра, консультации преподавателя.

Рекомендации по работе с литературой

Для более эффективного освоения и усвоения материала рекомендуется ознакомиться с теоретическим материалом по той или иной теме до проведения лабораторного занятия. Всю учебную литературу желательно изучать «под конспект». Цель написания конспекта по дисциплине – сформировать навыки по поиску, отбору, анализу и формулированию учебного материала. При работе над конспектом обязательно выявляются и отмечаются трудные для самостоятельного изучения вопросы, с которыми уместно обратиться к преподавателю при посещении консультаций, либо в индивидуальном порядке.

Подготовка к практическим занятиям

Подготовку к практической работе студент должен начать с изучения теоретического материала и ознакомления с планом, который отражает содержание предложенной темы. Все новые понятия по изучаемой теме необходимо выучить наизусть и внести в глоссарий, который целесообразно вести с самого начала изучения курса. Результат такой работы должен проявиться в способности студента свободно ответить на теоретические вопросы по теме задания, и правильном его выполнении.

В процессе выполнения практической работы студент должен создать требуемый документ с помощью предлагаемого программного средства и выполнить требуемые в задании операции. Задание содержит методические указания по подготовке документа, который должен быть получен в результате выполнения работы. При подготовке следует их внимательно прочесть.

Приложение 2

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

«Дальневосточный федеральный университет»

(ДВФУ)

ШКОЛА ЕСТЕСТВЕННЫХ НАУК

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

по дисциплине «Основы аналитики больших объемов данных»

Направление подготовки – 09.04.04 «Программная инженерия»

Магистерская программа «Разработка программно-информационных систем»

Форма подготовки (очная)

Владивосток

2018

Паспорт
фонда оценочных средств
по дисциплине «Основы аналитики больших объемов данных»

Код и формулировка компетенции	**Этапы формирования компетенции**	
ОК-4 умением быстро осваивать новые предметные области, выявлять противоречия, проблемы и вырабатывать альтернативные варианты их решения	Знает	Особенности существующих алгоритмов и технологий обработки данных.
	Умеет	Выявлять противоречия алгоритмов при их использовании для конкретных задач.
	Владеет	Методами адаптации алгоритмов для альтернативных решений.
ПК-2 знанием методов научных исследований и владением навыками их проведения	Знает	Методы поиска литературы по новым алгоритмам и технологиям обработки больших объемов данных.
	Умеет	Выделять в алгоритмах основное.
	Владеет	Методами сравнения алгоритмов.
ПК-4 владением существующими методами и алгоритмами решения задач распознавания и обработки данных	Знает	Основные алгоритмы решения задач распознавания и обработки данных.
	Умеет	Применять алгоритмы при анализе больших объемов данных.
	Владеет	Методами выбора подходящих алгоритмов для конкретных типов задач.

№ п/п	Контролируемые разделы/темы дисциплины	Коды и этапы формирования компетенций		Оценочные средства – наименование	
				текущий контроль	промежуточная аттестация
1	Темы 1-2	ОК-4 ПК-2 ПК-4	Знает	Собеседование УО-1, круглый стол УО-4	Зачет Вопросы 1-4
			Умеет	Практическое занятие 1 ПР-6	
2	Темы 3-4	ОК-4 ПК-2 ПК-4	Знает	Собеседование УО-1, круглый стол УО-4	Зачет Вопросы 5-7
			Умеет	Практическое занятие 2 ПР-6	
3	Темы 5-6	ОК-4 ПК-2 ПК-4	Знает	Собеседование УО-1, круглый стол УО-4	Зачет Вопросы 8-10
			Умеет	Практические занятия 3-5 ПР-6	

Шкала оценивания уровня сформированности компетенций

Код и формулировка компетенции	**Этапы формирования компетенции**		**критерии**	**показатели**
ОК-4 умение быстро осваивать новые предметные области, выявлять противоречия, проблемы и вырабатывать альтернативные варианты их решения	знает (пороговый уровень)	Особенности существующих алгоритмов и технологий обработки данных	Знание областей применения, преимуществ и недостатков основных алгоритмов и технологий обработки данных	Способность выбрать наиболее подходящий алгоритм и технологию обработки данных для произвольной предметной области.
	умеет (продвинутый)	Выявлять противоречия алгоритмов при их использовании для конкретных задач	Умение оценить степень применимости, а также временную и вычислительную сложности выбранного алгоритма для решения конкретной прикладной задачи.	Способность оценить временные и ресурсные затраты на решения конкретной прикладной задачи выбранным алгоритмом обработки данных.
	владеет (высокий)	Методами адаптации алгоритмов для альтернативных решений	Владение способностью провести анализ степени применимости имеющихся алгоритмов для всех возможных (альтернативных) решений конкретной прикладной задачи.	Способность выбрать из имеющегося набора алгоритмов одного, наиболее подходящего с точки зрения экономии ресурсов, а также качества получаемого результата.
ПК-2 знание методов научных исследований и владением навыками их проведения	знает (пороговый уровень)	Методы поиска литературы по новым алгоритмам и технологиям обработки больших объемов данных	Знание внешних и внутренних свойств алгоритмов обработки больших объемов данных, а также современных технологий в этой области.	Способность найти необходимую техническую литературу по новым алгоритмам и технологиям обработки больших объемов данных.
	умеет (продвинутый)	Выделять в алгоритмах основное	Умение сформулировать основные	Способность провести компьютерный

			требования к свойствам алгоритма.	эксперимент для вычисления оценок свойств алгоритмов.
	владеет (высокий)	Методами сравнения алгоритмов	Владение схемой компьютерного эксперимента для вычисления оценок свойств алгоритмов.	Способность сравнить два алгоритма на основе их внешних и внутренних оценок.
ПК-4 владение существующими методами и алгоритмами решения задач распознавания и обработки данных	знает (пороговый уровень)	основные алгоритмы решения задач распознавания и обработки данных	Знание областей применения, преимуществ и недостатков основных методов распознавания и обработки данных.	Способность выбрать наиболее подходящий метод распознавания и обработки данных для произвольной предметной области.
	умеет (продвинутый)	применять алгоритмы при анализе больших объемов данных	Умение реализовать заданный алгоритм и применить его к заданной обучающей выборке большого объема.	Способность подготовить набор данных и проанализировать его при помощи конкретного алгоритма.
	владеет (высокий)	методами выбора подходящих алгоритмов для конкретных типов задач	Владение навыками выделения и анализа совокупностей признаков из набора данных и подбора для них наиболее подходящего алгоритма обработки данных.	Способность выделить из группы признаков подгруппу, в наибольшей степени влияющей на результат, и в зависимости от нее способность подобрать наиболее подходящий алгоритм.

Методические рекомендации, определяющие процедуры оценивания результатов освоения дисциплины

Промежуточный контроль

Промежуточный контроль осуществляется в конце семестра и завершает изучение дисциплины. Помогает оценить более крупные совокупности знаний и умений, сформированность определенных профессиональных компетенций по дисциплине. Промежуточный контроль проводится в форме зачета, допуск к зачету возможен для обучающихся, получивших оценку «зачтено» в результате выполнения самостоятельной работы и успешно выполнившие все практические задания.

Оценочные средства для промежуточной аттестации

Вопросы к зачету

1. Что означает термин «Big Data» в информационных технологиях?
2. Что является основной целью обработки Big Data?
3. Кто и в каком году впервые ввел термин «Big Data»?
4. Какие главные характеристики Big Data?
5. Какие данные занимают больше мировой памяти относительно остальных?
6. Какие понятия содержит в себе принцип трех «V»?
7. С какого года Большие данные изучаются как академический предмет в вузовских программах по науке о данных?
8. Что является примером квази-структурированных данных?
9. Как назывался первый суперкомпьютер, оснащенный вопросно-ответной системой искусственного интеллекта?
10. Чем характеризуются «Большие данные»?
11. Что является главным результатом процесса Business Intelligence?
12. Что означает термин «Business Intelligence» в информационных технологиях?
13. Расшифруйте аббревиатуру OLAP.
14. Что относится к средствам предоставления информации в Business Intelligence?
15. Что относится к средствам интеграции в «Business Intelligence»?
16. Какие цели ставит перед собой Data Science?
17. Что такое жизненный цикл аналитики данных?
18. Дайте определение термину «предиктивное моделирование»?
19. Что такое ETL?
20. Какова роль BI-аналитика в проекте?
21. Что такое Apache Hadoop?
22. В чем преимущества решений на базе Hadoop?
23. Что такое MapReduce?
24. Какими достоинствами и недостатками обладает MapReduce?
25. Какому основному принципу следует HDFS?
26. Какой размер блока по умолчанию в HDFS?
27. Какие функции выполняет NameNode в HDFS?
28. Какой узел отвечает за репликацию данных в Hadoop?
29. Какие компоненты содержит Slave узел в Hadoop?
30. Какие компоненты содержит Master узел в Hadoop?
31. Какие компоненты являются частями HDFS?

32. Какое API было добавлено в Hadoop v2.0?
33. Для чего используется автономный режим Hadoop?
34. Какой режим необходим для того, чтобы на локальной машине использовать Hadoop как кластер, состоящий из одного узла?
35. Что является отличительной особенностью NoSQL?
36. В каком случае стоит применять NoSQL хранилища?
37. Что, согласно теореме CAP, возможно обеспечить в любой реализации распределённых вычислений?
38. Какое свойство означает, что транзакции не нарушают согласованность данных, то есть они переводят базу данных из одного корректного состояния в другое?
39. Какой способ хранения данных используется в MongoDB?
40. Что относится к плюсам репликации?
41. Что относится к преимуществам нереляционных БД?
42. На какие три группы подразделяют пользователей в MongoDB?
43. Что такое шардинг?
44. Какие три свойства фигурируют в определении теоремы CAP?
45. Для чего нужна визуализация?
46. Как называется один из самых популярных языков сценариев?
47. Какие достоинства у Amazon S3?
48. Какие традиционные виды визуализации?
49. Какие отличия и основные возможности у языка R?
50. В чем особенности хранения в Amazon S3?
51. Что такое дедупликация данных?
52. В чем основные задачи визуализации?
53. Какие требования предъявляются к визуализации?
54. Какие типы визуализации можно выделить?
55. Чем анализ больших данных отличается от традиционного анализа?
56. Какие основные типы Data Mining?
57. Какие категории Web Mining можно выделить?
58. В чем основная задача Web Content Mining?
59. В чем основные задачи интеллектуального анализа текстов?

Критерии выставления оценки магистранту на зачете

Баллы (рейтинговой оценки)	**Оценка** (стандартная)	**Требования к сформированным компетенциям**
86-100 76-85 61-75	«зачтено»	Оценка «зачтено» выставляется магистранту, если он: – глубоко и прочно усвоил программный материал, исчерпывающе, последовательно, четко и логически стройно его излагает, умеет тесно увязывать теорию с практикой, свободно справляется с задачами, вопросами и другими видами применения знаний, причем не затрудняется с ответом при видоизменении заданий, правильно обосновывает принятое решение, владеет разносторонними навыками и приемами выполнения практических задач; – твердо знает материал, грамотно и по существу излагает его, не допуская существенных неточностей в ответе на вопрос, правильно применяет теоретические положения при решении практических вопросов и задач, владеет необходимыми навыками и приемами их выполнения; – имеет знания только основного материала, но не усвоил его деталей, допускает неточности, недостаточно правильные формулировки, нарушения логической последовательности в изложении программного материала, испытывает затруднения при выполнении практических работ.
0-60	«незачтено»	Оценка «незачтено» выставляется магистранту, который не знает значительной части программного материала, допускает существенные ошибки, неуверенно, с большими затруднениями выполняет практические работы. Как правило, оценка «незачтено» ставится магистрантам, которые не могут продолжить обучение без дополнительных занятий по соответствующей дисциплине.

Критерии оценки проектов

- 100-86 баллов выставляется, если магистрант/группа точно определили содержание и составляющие части задания, умеют аргументировано отвечать на вопросы, связанные с заданием. Продемонстрировано знание и владение навыками самостоятельной исследовательской работы по теме. Фактических ошибок, связанных с пониманием проблемы, нет.
- 85-76 баллов – работа магистранта/группы характеризуется смысловой цельностью, связностью и последовательностью изложения; допущено не более 1 ошибки при объяснении смысла или содержания

проблемы. Продемонстрированы исследовательские умения и навыки. Фактических ошибок, связанных с пониманием проблемы, нет.

- 75-61 балл – проведен достаточно самостоятельный анализ основных этапов и смысловых составляющих проблемы; понимание базовых основ и теоретического обоснования выбранной темы. Привлечены основные источники по рассматриваемой теме. Допущено не более 2 ошибок в смысле или содержании проблемы.
- 60-50 баллов – если работа представляет собой пересказанный или полностью переписанный исходный текст без каких бы то ни было комментариев, анализа. Не раскрыта структура и теоретическая составляющая темы. Допущено три или более трех ошибок смыслового содержание раскрываемой проблемы.

Шкала оценивания проектов

Менее 60 баллов	Не зачтено
От 61 до 75 баллов	зачтено
От 76 до 85 баллов	зачтено
От 86 до 100 балов	зачтено

Текущий контроль

Текущий контроль предполагает систематическую проверку усвоения учебного материала, сформированности компетенций или их элементов, регулярно осуществляемую на протяжении изучения дисциплины, в соответствии с ее рабочей программой.

Состоит в проверке правильности выполнения заданий по самостоятельной работе. Задание зачтено, если нет ошибок. По текущим ошибкам даются пояснения.

Тесты предназначены для проверки знаний по компетенциям. Проверка достижения умений и навыков по компетенциям проверяется выполнением практических работ и курсовой работы.

Примерные тесты для проверки сформированности компетенций

ОК-4 умением быстро осваивать новые предметные области, выявлять противоречия, проблемы и вырабатывать альтернативные варианты их решения	Знает особенности существующих алгоритмов и технологий обработки данных.

1. В принцип «Трёх V» не входит один из перечисленных признаков:	Ответы: **а. Veracity (достоверность данных): в настоящее время достоверность имеющихся данных является важнейшим критерием для пользователей. Недостоверная информация приводит к затруднению анализа данных.** б. Volume (объем): накопленная база данных представляет собой гигантский объем информации, для которого обработка и хранение традиционными способами являются трудоёмкими процессами. Такой объем нуждается в новых подходах и в более усовершенствованных инструментах. в. Variety (многообразие): данная характеристика означает возможность одновременной обработки структурированной и неструктурированной информации различных форматов. Главным отличием структурированной информации является возможность классификации.
2. К полуструктурированным данным относятся:	Ответы: а. Данные, которые не имеют определённой формы, могут включать в себя видео, аудио файлы, свободный текст, информацию, поступающую из социальных сетей. **б. Данные, которые не соответствуют чёткой структуре таблиц и отношений в реляционных базах данных, однако такие данные содержат специальные теги и иные маркёры, позволяющие отделить семантические элементы.** в. Данные, определяющие конкретную предметную область, упорядоченные специальным образом и организованные таким образом, чтобы над ними можно было выполнить анализ.

ПК-2 знанием методов научных исследований и владением навыками их проведения	Знает методы поиска литературы по новым алгоритмам и технологиям обработки больших объемов данных.
1. Предиктивное моделирование (Predictive Modelling) – это:	**а. Процесс создания (или выбора) модели для предсказания вероятности наступления некоторого события.** б. Компьютерная техника извлечения знаний, которая использует искусственный интеллект для распознавания образов и выделения значимых закономерностей из данных, находящихся в хранилищах или входных или выходных потоках.

	в. Методы и инструменты для перевода необработанной информации в осмысленную, удобную форму.
2. Text Mining – это:	Ответы: **а. Нетривиальный процесс обнаружения действительно новых, потенциально полезных и понятных шаблонов в неструктурированных текстовых данных.** б. Использование методов интеллектуального анализа для автоматического обнаружения веб-документов и услуг, извлечения информации из веб-ресурсов и сервисов. в. Собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

ПК-4 владением существующими методами и алгоритмами решения задач распознавания и обработки данных	Знает основные алгоритмы решения задач распознавания и обработки данных.
1. Достоинством модели Map Reduce не является:	Ответы: а. Автоматическое распараллеливание (функции Map и Reduce могут выполняться параллельно и независимо друг от друга) б. Масштабируемость (данные могут располагаться и обрабатываться на разных серверах). **в. Фиксированный алгоритм обработки данных.** г. Отказоустойчивость (при отказе сервера функции Map и Reduce запускаются на другом сервере).
2. Ошибка 2-ого рода – это:	Ответы: **а. «Ложное обнаружение», когда при отсутствии события ошибочно выносится решение о его присутствии.** б. «Ложный пропуск», когда интересующее событие ошибочно не обнаруживается.