Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Сибирский государственный индустриальный университет»

Кафедра прикладных информационных технологий и программирования

УТВЕРЖДАЮ
Директор института
информационных технологий и
автоматизированных систем
_____ А.В. Зимин
подпись
«____» _____ 20___ г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Наука о данных и аналитика больших данных

09.04.03 «Прикладная информатика»
(направленность (профиль): «Искусственный интеллект и машинное обучение»)

Квалификация выпускника
Магистр

Форма обучения
Очная форма

Срок обучения: 2 года

Год начала подготовки 2026

Новокузнецк
2026

СибГИУ
Подлинник электронного документа хранится в системе электронного согласования документов СибГИУ
ДОКУМЕНТ ПОДПИСАН ЭП
Сертификат: daba56d7b892efca3bf5c04ed0251c98b19b099c
Владелец: Директор ИИТиАС Зимин Алексей Валерьевич
Действителен: с 15 января 2025 г. по 10 апреля 2026 г.
Дата и время: 23-02-2026 19:30:08 ПРОВЕРИТЬ ПОДПИСЬ

1 Цели и задачи освоения учебной дисциплины

Целями учебной дисциплины являются:

– овладение методами анализа и обработки больших объемов данных для принятия обоснованных решений на основе современных технологий.

Задачами учебной дисциплины являются:

– Изучение основных теоретических принципов обработки данных, включая методы сбора, хранения, анализа и визуализации информации;

– овладение практическими навыками работы с различными инструментами и технологиями для обработки и анализа больших объемов данных, такими как Python, SQL, алгоритмы машинного обучения и фреймворки для работы с Big Data.;

– практическое применение полученных знаний и навыков на практике через выполнение проектов по обработке и анализу данных.

2 Место учебной дисциплины в структуре ООП по направлению подготовки (специальности)

Учебная дисциплина относится к учебным дисциплинам части, формируемой участниками образовательных отношений **Блока 1 «Дисциплины (модули)»** ООП по направлению подготовки (специальности) 09.04.03 «Прикладная информатика».

Учебная дисциплина базируется на предварительном усвоении обучающимися учебных дисциплин:

– Математические и инструментальные методы анализа данных;

– Машинное обучение;

– Программирование глубоких нейронных сетей на Python;

– Анализ текстовых данных.

Учебная дисциплина дополняет знания и умения, получаемые по одновременно изучаемым и последующим учебным дисциплинам:

– Методы и инструменты цифровой трансформации.

3 Планируемые результаты обучения по учебной дисциплине

Процесс изучения учебной дисциплины направлен на формирование следующих компетенций:

– Профессиональные компетенции

Наименование категории (группы) ПК	Код и наименование ПК	Код и наименование индикатора достижения ПК	Планируемые результаты обучения
	ПК-1: Способен участвовать в	ПК-1.1 Принимает участие в разработке моделей данных,	– знать: методы моделирования данных, включая

	управлении работами по получению, хранению и обработки больших объемов данных	проводит анализ больших объемов данных, строит модели на основе данных	статистическое, машинное обучение и подходы к глубокому обучению. – уметь: разрабатывать модели данных для конкретных приложений, анализировать большие объемы данных, используя соответствующие инструменты и методы.
		ПК-1.2 Принимает участие в разработке методов преобразования данных	– знать: основы различных методов преобразования данных, включая очистку данных, нормализацию, агрегацию и инженерию функций. – уметь: применять инструменты и методы для эффективного преобразования необработанных данных.
		ПК-1.3 Принимает участие в разработке методов обработки больших данных	– знать: методы обработки больших данных, включая распределенные вычисления, параллельную обработку и алгоритмы интеллектуального анализа данных. – уметь: применять алгоритмы и методы программирования для разработки и реализации эффективных и масштабируемых решений.

4 Объем и содержание учебной дисциплины

Учебные занятия по учебной дисциплине проводятся в форме контактной работы и в форме самостоятельной работы обучающихся.

Контактная работа включает в себя занятия лекционного типа (лекции), занятия семинарского типа (семинары, практические занятия,

практикумы), промежуточную аттестацию обучающихся и иные формы взаимодействия обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации ООП на иных условиях, в том числе при проведении промежуточной аттестации обучающихся. Контактная работа может проводиться с применением электронного обучения, дистанционных образовательных технологий.

Объем учебной дисциплины

Семестр / курс		***ИТОГО***	***3 семестр***
Форма промежуточной аттестации			*экзамен*
Трудоёмкость	*академ. час.*	***108***	*108*
	зачетных единиц	***3***	*3*
Лекции, *академ. час.*		***8***	*8*
в форме практической подготовки		***0***	*0*
Лабораторные работы, *академ. час.*		***0***	*0*
в форме практической подготовки		***0***	*0*
Практические занятия, *академ. час.*		***8***	*8*
в форме практической подготовки		***0***	*0*
Курсовая работа / проект, *академ. час.*		***0***	*0*
в форме практической подготовки		***0***	*0*
Консультации, *академ. час.*		***0***	*0*
в форме практической подготовки		***0***	*0*
Самостоятельная работа, *академ. час.*		***56***	*56*
в форме практической подготовки		***0***	*0*
Контроль, *академ. час.*		***36***	*36*
в форме практической подготовки		***0***	*0*

Содержание учебной дисциплины

Раздел 1 Введение в большие данные;

Тема 1.1 Определение больших данных и причины их появления. Примеры возможностей для бизнеса;

Тема 1.2 Различие между Business Intelligence и Big Data;

Раздел 2 Жизненный цикл аналитики данных;

Тема 2.1 Понятие жизненного цикла аналитики данных;

Тема 2.2 Роли, необходимые для успешного создания проекта по аналитике данных;

Раздел 3 Высокопроизводительные вычисления;

Тема 3.1 Распределенные вычисления на нескольких серверах, вычислительная парадигма MapReduce;

Тема 3.2 Проект Apache Hadoop и его экосистема. Apache Spark и его компоненты. Вычисления в реальном времени, Apache Storm, Flink;

Раздел 4 Масштабирование и многоуровневое хранение данных;

Тема 4.1 Теорема CAP;

Тема 4.2 Парадигма NoSQL. Классификация NoSQL баз данных;

Раздел 5 Визуализация данных и результатов анализа;

Тема 5.1 Техники визуализации данных, введение в язык R;

Тема 5.2 Визуализация данных в R;

Раздел 6 Сложные методы аналитики;

Тема 6.1 Классификация задач анализа: Text, Data, Web, Social Mining;

Тема 6.2 Применение машинного обучения в аналитике. K-means и C-means кластеризация, классификация. Логистическая регрессия, ассоциации, алгоритм Априори;

Раздел 7 Анализ текста;

Тема 7.1 Поисковые механизмы: Lucene, Solr, ElasticSearch;

Тема 7.2 Алгоритмы Work2Vec и Glove.

5 Перечень тем лекций

№ раздела / темы дисциплины	Темы лекций	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
Раздел 1.	Введение в большие данные	2	
Раздел 2.	Жизненный цикл аналитики данных	1	
Раздел 3.	Высокопроизводительные вычисления	1	
Раздел 4.	Масштабирование и многоуровневое хранение данных	1	
Раздел 5.	Визуализация данных и результатов анализа	1	
Раздел 6.	Сложные методы аналитики	1	
Раздел 7.	Анализ текста	1	
Итого:		**8**	**0**

6 Перечень тем практических занятий (семинаров)

№ раздела / темы дисциплины	Темы практических занятий (семинаров)	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
Раздел 1.	Введение в большие данные	2	
Раздел 2.	Жизненный цикл аналитики данных	1	
Раздел 3.	Высокопроизводительные вычисления	1	
Раздел 4.	Масштабирование и многоуровневое хранение данных	1	
Раздел 5.	Визуализация данных и результатов анализа	1	

Раздел 6.	Сложные методы аналитики	1	
Раздел 7.	Анализ текста	1	
Итого:		**8**	**0**

7 Перечень тем лабораторных работ

№ раздела / темы дисциплины	Темы лабораторных работ	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
	Отсутствуют		
Итого:		**0**	**0**

8 Перечень тем курсовых работ (проектов)

№ раздела / темы дисциплины	Темы курсовых работ (проектов)	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
	Отсутствуют		
Итого:		**0**	**0**

9 Виды самостоятельной работы

№ раздела / темы дисциплины	Виды самостоятельной работы	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
Раздел 1.	1. Изучение лекционного материала; 2. Подготовка к практическому занятию; 3. Прохождение тестирования.	10	
Раздел 2.	1. Изучение лекционного материала; 2. Подготовка к практическому занятию; 3. Прохождение тестирования.	5	
Раздел 3.	1. Изучение лекционного материала; 2. Подготовка к практическому занятию; 3. Прохождение тестирования.	5	
Раздел 4.	1. Изучение лекционного материала; 2. Подготовка к практическому занятию; 3. Прохождение тестирования.	6	

Раздел 5.	1. Изучение лекционного материала; 2. Подготовка к практическому занятию; 3. Прохождение тестирования.	10	
Раздел 6.	1. Изучение лекционного материала; 2. Подготовка к практическому занятию; 3. Прохождение тестирования.	10	
Раздел 7.	1. Изучение лекционного материала; 2. Подготовка к практическому занятию; 3. Прохождение тестирования.	10	
Контроль	*Подготовка к экзамену*	36	
Итого:		**92**	**0**

10 Матрица компетентностных задач по учебной дисциплине

Контролируемые разделы (темы) дисциплины	Контролируемые компетенции	Оценочные средства
Раздел 1.	ПК-1	1. Наблюдение за работой на практическом занятии; 2. Тестовые задания
Раздел 2.	ПК-1	1. Наблюдение за работой на практическом занятии; 2. Тестовые задания
Раздел 3.	ПК-1	1. Наблюдение за работой на практическом занятии; 2. Тестовые задания
Раздел 4.	ПК-1	1. Наблюдение за работой на практическом занятии; 2. Тестовые задания
Раздел 5.	ПК-1	1. Наблюдение за работой на практическом занятии; 2. Тестовые задания
Раздел 6.	ПК-1	1. Наблюдение за работой на практическом занятии; 2. Тестовые задания
Раздел 7.	ПК-1	1. Наблюдение за работой на практическом занятии; 2. Оценка участия в работе на лекционном занятии; 3. Тестовые задания
Экзамен.	ПК-1	1. Экзамен на основе результатов компьютерного тестирования;

		2. Экзамен на основе результатов текущего контроля

11 Входной контроль

11.1 Цель входного контроля − определить начальный уровень подготовленности обучающихся и выстроить индивидуальную траекторию обучения. В условиях личностно-ориентированной образовательной среды результаты входного оценивания обучающегося используются как начальные значения в индивидуальном профиле академической успешности обучающегося.

11.2 Описание оценочных средств

Форма проведения входного контроля – тестирование.

Длительность тестирования – 30 минут.

Тестирование проводится в течение первых двух недель изучения учебной дисциплины в СУО СибГИУ.

Для проведения тестирования используются тестовые задания для входного контроля (для ответа на 1 вопрос, как правило, выделяется 1 минута).

Шкала оценивания результатов тестирования

За каждый правильный ответ выставляется 1 балл. Оценка выставляется в зависимости от процента правильных ответов:

Оценка формируется в соответствии с критериями таблицы:

Максимальный балл	Оценка
10	зачтено
4	Не зачтено

Основываясь на результатах тестирования, формируется индивидуальная траектория обучающегося (группы обучающихся), исходя из следующих соображений:

– 5 – 10 – обучающийся готов к изучению учебной дисциплины без дополнительного повторения изученного ранее материала;

– 4 и менее – обучающийся имеет слабую подготовку к изучению учебной дисциплины, необходимы корректирующие занятия и консультации.

Тестовые задания для входного контроля знаний обучающихся приведены в приложении.

12 Текущий контроль

12.1 Текущий контроль знаний используется для оперативного и регулярного управления учебной деятельностью (в том числе

самостоятельной) обучающихся. Данный вид контроля стимулирует у обучающихся стремление к систематической самостоятельной работе по изучению учебной дисциплины и позволяет оценить степень освоения изученного материала.

Текущий контроль успеваемости осуществляется в течение семестра, в ходе повседневной учебной работы с ис-пользованием онлайн-курса Наука о данных и аналитика больших данных, доступного через сеть Интернет, размещенного по адресу: https://openedu.ru/course/spbstu/BIGDATA/.

Формы проведения текущего контроля – наблюдение за работой на практическом занятии (оценка результатов выполнения в онлайн-курсе; тестовые задания (в онлайн-курсе Наука о данных и аналитика больших данных).

В процессе обучения осуществляется тьюторское (организационно-методическое) и техническое сопровождение самостоятельной работы слушателей в онлайн-курсе, предоставление условий для прохождения слушателем контрольного испытания с идентификации личности слушателя в процессе контроля, с целью проверки результатов освоения онлайн-курса.

Самостоятельная работа в онлайн-курсе является асинхронной и осуществляется с использованием инструментов электронно-информационной образовательной среды, интернет-сервисов и дистанционного онлайн-взаимодействия преподавателя (тьютора) он-лайн-курса.

Ход обучения и результаты текущего и промежуточного контроля знаний фиксируются в электронной ведомости обучающегося, что позволяет контролировать учебные достижения обучающегося и своевременно проводить коррекцию его обучения.

В рамках текущего контроля формируется оценка по стобалльной шкале. При этом вклад в итоговую оценку по курсу распределяется следующим образом:

Аттестация – 50%.

Итоговый экзамен – 50%.

В онлайн-курсе задан вес оценок.

Взвешенная оценка рассчитывается умножением оценки на вес задания. Взвешенная оценка используется для расчета балла за курс. Сводка по оценкам

Тип задания	Вес	Оценка	Проходной балл	Взвешенная оценка
Аттестация	50%	x	50%	0,5*x
Итоговый экзамен	50%	y	50%	0,5*y

12.2 Описание оценочных средств

Шкалы оценивания (методики оценки)

12.2.1 Рекомендации по оцениванию результатов тестирования обучающихся

Тестирование проводится по изученным разделам (темам) учебной дисциплины для выявления уровня знаний и умений обучающегося, мотивирования обучающегося к активизации самостоятельной работы по усвоению учебного материала, выявления и устранения пробелов в знаниях, повышения дисциплины и организации учебной деятельности обучающегося.

Тестирование проводится на платформе «Открытое образование» в онлайн-курсе Наука о данных и аналитика больших данных на Python в рамках часов, отведенных на самостоятельную работу обучающихся и частично, отведенные на выполнение практических заданий. В курсе предусмотрены тесты самоконтроля, контроль по модулям курса и итоговый контрольный тест.

12.2.2 Рекомендации по оцениванию в процессе наблюдения за работой на практическом занятии

Наблюдение за работой на практическом занятии – важнейший компонент контроля деятельности обучающихся и качества их обучения.

Критерии оценки в процессе наблюдения за работой на практическом занятии:

– самостоятельность выполнения задания и работы с методическими указаниями (учитывается работа в течение занятия, быстрота и способность нахождения в методических указаниях нужной информации);

– правильность выполнения работы (учитывается последовательность выполняемых действий, умение работать с инструментами и/или средствами вычислительной техники);

– рациональность использования времени, отведенного на задание (не одобряется затянутость выполнения работы с учетом индивидуальных особенностей обучающихся);

– активность обучающегося при выполнении практической работы (учитывается индивидуальная работа и работа в группе).

Критерии оценки по результатам наблюдения за работой на практическом занятии

Наименование параметра оценки	Максимальное количество баллов
Самостоятельно выполняет задания; умеет работать с методическими указаниями.	2,5
Указанные в методических указаниях действия выполняет в правильной последовательности; умеет работать с инструментами и/или средствами вычислительной техники.	1,25
Рационально использует время, отведенное на задание; проявляет активность при выполнении практической работы.	1,25

Наименование параметра оценки	Максимальное количество баллов
Итого	**5**

Максимальное количество – 5 баллов соответствует оценке «отлично», 4 балла – «хорошо», 3 балла – «удовлетворительно», 2 балла – «неудовлетворительно».

Задания и вопросы для текущего контроля знаний обучающихся по разделам (темам) учебной дисциплины приведены в приложении.

13 Промежуточная аттестация

13.1 Оценочные материалы для промежуточной аттестации обучающихся по учебной дисциплине предназначены для оценки степени достижения запланированных результатов обучения по завершению изучения учебной дисциплины в установленной учебным планом форме и позволяют определить качество усвоения изученного материала.

Итоговой формой контроля сформированности компетенций у обучающихся по учебной дисциплине является: экзамен.

ФОС промежуточной аттестации включает: экзамен на основе результатов компьютерного тестирования; экзамен на основе результатов текущего контроля. Промежуточная аттестация производится на основе текущего контроля успеваемости при условии выполнения всех заданий в семестре, предусмотренных фондом оценочных средств по дисциплине, и выполнения итогового контрольного теста с прокторингом на платформе «Открытое образование» в онлайн-курсе "Наука о данных и аналитика больших данных". – (https://openedu.ru/course/spbstu/BIGDATA/).

13.2 Оценивание обучающегося на экзамене

Оценка	Требования к знаниям
отлично	Оценка «отлично» выставляется обучающемуся, если он глубоко и прочно усвоил программный материал, исчерпывающе, последовательно, четко и логически 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, использует в ответе материал различной литературы, правильно обосновывает принятое нестандартное решение, обладает приемами выполнения практических задач по формированию компетенций.
хорошо	Оценка «хорошо» выставляется обучающемуся, если он твердо знает материал, грамотно и по существу излагает его, не допуская существенных неточностей в ответе на вопрос, правильно применяет теоретические положения при решении

Оценка	Требования к знаниям
	практических вопросов и задач, обладает необходимыми умениями и приемами их выполнения.
удовлетворительно	Оценка «удовлетворительно» выставляется обучающемуся, если он имеет знания только основного материала, но не усвоил его деталей, допускает неточности, недостаточно правильные формулировки, испытывает сложности при выполнении практических работ и затрудняется связать теорию вопроса с практикой.
неудовлетворительно	Оценка «неудовлетворительно» выставляется обучающемуся, который не знает значительной части программного материала, неуверенно отвечает, допускает серьезные ошибки, не имеет представлений по методике выполнения практической работы. Как правило, оценка «неудовлетворительно» ставится обучающимся, которые не могут продолжить обучение без дополнительных занятий по данной учебной дисциплине.

Задания и вопросы для промежуточной аттестации обучающихся приведены в приложении.

14 Учебно-методическое и информационное обеспечение учебной дисциплины

а) литература:

1 Рашка, С. Python и машинное обучение практическое пособие / С. Рашка; пер. с англ. А. В. Логунова. - Москва : ДМК Пресс, 2017. - 418 с. - ISBN 978-5-97060-409-0. – URL: https://www.studentlibrary.ru/book/ISBN9785970604090.html (дата обращения: 17.02.2026);

2 Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах; пер. с англ. А. А. Слинкина. - 2-е изд. - Москва : ДМК Пресс, 2023. - 401 с. - ISBN 978-5-89818-300-4. – URL: https://www.studentlibrary.ru/ru/book/ISBN9785898183004.html (дата обращения: 17.02.2026);

3 Мастицкий, С. Э. Статистический анализ и визуализация данных с помощью R / С. Э. Мастицкий, В. К. Шитиков. - 2-е изд. - Москва : ДМК Пресс, 2023. - 497 с. - ISBN 978-5-89818-601-2. – URL: https://www.studentlibrary.ru/ru/book/ISBN9785898186012.html (дата обращения: 17.02.2026).

б) ресурсы информационно-телекоммуникационной сети «Интернет»:

1 Электронный каталог : сайт / Научно-техническая библиотека СибГИУ. – Новокузнецк, [199 –]. – URL: http://libr.sibsiu.ru. – URL: https://libr.sibsiu.ru;

2 Электронная библиотека // Научно-техническая библиотека СибГИУ : сайт. – Новокузнецк, [200 –]. – URL: http://library.sibsiu.ru/LibrELibraryFullText.asp. – Режим доступа: для авторизир. пользователей. – URL: https://library.sibsiu.ru/LibrELibraryFullText.asp;

3 Консультант студента : электронно-библиотечная система / ООО «КОНСУЛЬТАНТ СТУДЕНТА». – Москва, [200 –]. – URL: http://www.studentlibrary.ru. – Режим доступа: для авторизир. пользователей;

4 ЛАНЬ : электронно-библиотечная система : [коллекция «Единая профессиональная база знаний для технических вузов - Издательство Лань" ЭБС»] / ООО «Издательство ЛАНЬ». – Санкт-Петербург, [200 –]. – URL: http://e.lanbook.com. – Режим доступа: для авторизир. пользователей;

5 Образовательная платформа ЮРАЙТ / ООО «Электронное издательство ЮРАЙТ». – Москва, [200 –]. – URL: https://urait.ru. – Режим доступа: для авторизир. пользователей;

6 Университетская библиотека онлайн : электронно-библиотечная система / ООО «Директ-Медиа». – Москва, [200 –]. – URL: https://biblioclub.ru. – Режим доступа: для авторизир. пользователей. – URL: http://www.biblioclub.ru;

7 НАУЧНАЯ ЭЛЕКТРОННАЯ БИБЛИОТЕКА eLIBRARY.RU : база данных / ООО «НЭБ». – Москва, [200 –]. – URL: http://elibrary.ru. – Режим доступа: по подписке;

8 Электронные периодические издания ИВИС : универсальная база данных / ООО «ИВИС». – Москва, [200 –]. – URL: http://eivis.ru. – Режим доступа: по подписке;

9 Онлайн курс «Наука о данных и аналитика больших объемов данных». – URL: https://openedu.ru/course/spbstu/BIGDATA/.

в) лицензионное и свободно распространяемое программное обеспечение:

– 7-Zip;
– Adobe Acrobat Reader;
– Astra Linux Special Edition;
– Kaspersky Endpoint Security;
– Microsoft Office;
– Microsoft Windows;
– OnlyOffice;
– Р7-Офис.

г) базы данных и информационно-справочные системы:

1 ГАРАНТ : справочно-правовая система / ООО «Правовой центр «Гарант». – Кемерово, [200 –]. – Режим доступа: компьютерная сеть Сиб. гос. индустр. ун-та.;

2 КонсультантПлюс : справочно-правовая система / ООО «Информационный центр АНВИК». – Новокузнецк, [199 –]. – Режим доступа: компьютерная сеть библиотеки Сиб. гос. индустр. ун-та.;

3 Техэксперт : информационно-справочная система / ООО «Группа компаний «Кодекс». – Кемерово, [200 –]. – Режим доступа: компьютерная сеть Сиб. гос. индустр. ун-та.

15 Материально-техническое обеспечение учебной дисциплины

Материально-техническое обеспечение учебной дисциплины включает учебные аудитории, оснащенные оборудованием, компьютерной техникой, и техническими средствами обучения, в том числе:

- учебную аудиторию для проведения занятий лекционного типа, оборудованную учебной доской, экраном и мультимедийным проектором;
- учебную аудиторию для проведения занятий семинарского типа (практических занятий), оснащенную учебной доской, экраном и мультимедийным проектором;
- учебную аудиторию (помещения) для групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации;
- помещения для самостоятельной работы, оснащенные компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду, научно-техническую библиотеку СибГИУ.

Обучение по дисциплине проводится с использованием онлайн-курса Наука о данных и аналитика больших объемов данных, доступного через сеть Интернет, размещенного по адресу: https://openedu.ru/course/spbstu/BIGDATA/. В процессе обучения осуществляется тьюторское (организационно-методическое) и техническое сопровождение самостоятельной работы слушателей в онлайн-курсе, предоставление условий для прохождения слушателем контрольного испытания с идентификации личности слушателя в процессе контроля, с целью проверки результатов освоения онлайн-курса. Самостоятельная работа в онлайнкурсе является асинхронной и осуществляется с использованием инструментов электронноинформационной образовательной среды, интернет-сервисов и дистанционного онлайн-взаимодействия преподавателя (тьютора) под руководством онлайн-курса. Ход обучения и результаты промежуточного и итогового контроля знаний фиксируются в электронной ведомости обучающегося, что позволяет контролировать учебные достижения обучающегося и своевременно проводить коррекцию его обучения. Содержание программы разработано

работниками Высшей школы программной инженерии СПбГУ: Дробинцев П.В., к.т.н., Воинов Н.В., к.т.н., Никифоров И.В., к.т.н.

Рабочая программа дисциплины составлена в соответствии с требованиями ФГОС ВО по направлению подготовки (специальности) 09.04.03 «Прикладная информатика».

Составитель(и):

доцент Маслова Елена Владимировна (кафедра прикладных информационных технологий и программирования).

Рабочая программа дисциплины рассмотрена и утверждена на заседании кафедры.

Согласована:

Системный аналитик группы
системной архитектуры департамента
информационных технологий
ООО "СИ СОФТ" Е.А. Шапошникова

Приложение

Тестовые задания для входного контроля знаний обучающихся

Выберите один правильный ответ.

1 Алфавит языка программирования – это…
a. набор символов русского языка, используемых при составлении любых текстов, называемых алфавитом языка
b. набор цифр и знаков препинания, используемых при составлении текстов, называемых алфавитом языка
c. разрешенный к использованию набор символов, с помощью которого могут быть образованы слова и величины данного языка
2 Переменная характеризуется
a. именем, типом, значением
b. именем, значением
c. значением, типом
3 Величина, значение которой меняется в процессе выполнения программы, называется
a. константой
b. переменной
c. меткой
4 Операции DIV и MOD допустимы над величинами
a. целочисленного типа
b. вещественного типа
c. символьного типа
5 Какое значение примет переменная X после выполнения оператора X=17 mod 4?
a. 1
b. 4
c. 4.25
6 Дано трехзначное число x=abc. Операция x div 10 mod 10 определяет
a. единицы
b. десятки
c. сотни
7 Цикл FOR используется в том случае, если
a. необходимо, чтобы тело цикла выполнялось хотя бы один раз
b. точно знаете, сколько раз должно быть выполнено тело цикла
c. хотите, чтобы проверка была произведена прежде, чем выполнено тело цикла
8 Массив - это
a. совокупность конечного числа данных различных типов
b. совокупность конечного числа данных одного типа
c. скалярный тип данных
9 Индекс массива определяет

a. положение элементов массива относительно друг друга
b. номер элемента в массиве
c. количество элементов
10 Разветвляющейся называется алгоритмическая конструкция,…
a. в которой выбирается один из нескольких возможных путей (вариантов) вычислительного процесса;
b. в которой все этапы решения задачи выполняются строго последовательно
c. организующая многократное выполнение указанного действия

Задания и вопросы для текущего контроля знаний обучающихся по разделам (темам) учебной дисциплины

2 Тестирование по результатам освоения лекционного материала и самостоятельной работы

Тестирование проводится для определения степени усвоения материалов лекций, а также выполнения самостоятельной работы в рамках часов, отведенных на самостоятельную работу обучающихся, в СУО «Moodle».

Перечень вопросов для тестирования:

Раздел 1 Введение в большие данные

1 Дайте определение Big Data:

a) комплексный набор инструментов обработки структурированных данных колоссальных объемов

b) комплексный набор методов обработки неструктурированных данных колоссальных объемов

c) комплексный набор подходов, инструментов и методов обработки структурированных и неструктурированных данных колоссальных объемов

d) комплексный набор методов обработки структурированных данных колоссальных объемов

2 Какие данные занимают больше мировой памяти относительно остальных?

a) Structured Data
b) Unstructured Data
c) Semi-Structured Data
d) Quasi-Structured Data

3 Кто ввел термин Большие данные?

a) Клиффорд Линч
b) Алан Тьюринг
c) Бьерн Страуструп
d) Дональд Кнут

4 Какой из ниже перечисленных принципов работы не применятся к Big Data?

a) Горизонтальная масштабируемость

b) Вертикальная масштабируемость

c) Отказоустойчивость

d) Локальность данных

5 Выберите правильный порядок перечисления пяти фаз, составляющих процесс развития любой новой технологии.

a) Появление технологии, завышенные ожидания, разочарование, просветление, реализация;

b) Появление технологии, просветление, реализация, завышенные ожидания, разочарование;

c) Появление технологии, реализация, просветление, исправление, успех;

d) Правильного ответа нет.

6 Что не относится к неструктурированной информации?

a) Видео-контент

b) Аудио-контент

c) Текстовые документы

d) XML – документы

7 Какая компания создала технологию MapReduce?

a) Google

b) Yahoo

c) EMC

d) Oracle

8 Какая из характеристик не является основной для «Больших данных».

a) Скорость (Velocity)

b) Объём (Volume)

c) Многообразие (Variety)

d) Качество (Quality)

9 Основной целью обработки Big Data является

a) Получение воспринимаемых человеком результатов

b) Получение воспринимаемых компьютером результатов

c) Извлечение знаний

d) Структурирование знаний

10 Чему примерно равен объем всей существующей на земле информации (в байтах)?

a) 10^11

b) 10^21

c) 10^1010101

d) 10^171

Раздел 2 Жизненный цикл аналитики данных

1 Что означает термин «Business Intelligence» в информационных технологиях?

a) Комплексный набор методов для создания бизнес планов.

b) Методы и инструменты для перевода необработанной информации в осмысленную, удобную для восприятия форму.

c) Файлы, содержащие информацию о бизнес плане.

d) Технологии, направленные на развитие бизнеса.

2 К каким средствам «Business Intelligence» относится Интеграция с MS Office?

a) Средства предоставления информации

b) Средства интаграции

c) Средства анализа

d) Ко всем средствам

3 Что не является целью процесса Business Intelligence?

a) Интерпретация большого количества данных;

b) Моделирование исходов различных вариантов действий;

c) Модификация существующего программного обеспечения;

d) Отслеживание результатов решений.

4 Что из перечисленного не является средством анализа?

a) Продвинутая визуализация

b) Reporting

c) Predictive Modelling

d) Data Mining

5 Что такое хранилище данных?

a) Предметно-ориентированная информационная база данных, специально разработанная и предназначенная для подготовки отчётов и бизнес-анализа.

b) Носитель информации или удаленный сервер, способные хранить, получать, отдавать и обрабатывать данные.

c) Система из информационных хранилищ, предназначенных для хранения информации обрабатываемой кластером.

d) Все выше перечисленное.

6 Что относится к средствам предоставления информации в «Business Intelligence»?

a) Генератор нерегламентированных запросов

b) Совместная работа и управление рабочими процессами

c) Предиктивное моделирование и Data Mining

d) Карты показателей

7 Дайте определение процессу "Интеллектуальный анализ данных"?

a) Методы и инструменты для перевода необработанной информации в осмысленную, удобную форму

b) Процесс, ориентированный на бизнес-пользователя и включающий доступ и исследование информации, ее анализ, выработку интуиции и понимания, которые ведут к улучшенному и неформальному принятию решений

c) 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

d) Процесс создания (или выбора) модели для предсказания вероятности наступления некоторого события

8 Какой из этапов не относится к Business Intelligence?

a) Техническая обработка данных разных форматов;

b) Создание бизнес-отчётов по результатам анализа;

c) Прогнозирование;

d) OLAP.

9 Интеллектуальный анализ данных - это...

a) Предподготовка данных человеком для дальнейшего исследования машиной.

b) Искусственный интеллект способный распознавать образы и выделять значимые закономерности из данных.

c) Компьютерная техника извлечения знаний.

d) Множественное сканирование данных для выделения последовательностей.

10 Что из нижеперечисленного не является типом инструментальных средств «Business Intelligence»?

a) Средства предоставления информации

b) Средства интеграции

c) Средства анализа

d) Средства сжатия

Раздел 3 Высокопроизводительные вычисления

1 Что такое Spark?

a) Это библиотека для манипуляций над графами и выполнения с ними параллельных операций

b) Это базовый движок для крупномасштабной параллельной и распределенной обработки данных

c) Это проект Apache, который позиционируется как инструмент для «молниеносных кластерных вычислений»

d) Нереляционная распределенная база данных, исполняемая на верху HDFS

2 Дайте определение Map Reduce

a) Модель распределенных вычислений, предназначенная для параллельных вычислений над очень большими (до нескольких петабайт) объемами данных

b) Набор компонентов и интерфейсов для распределенных файловых систем и общего ввода-вывода

c) Распределенная файловая система, работающая на больших кластерах типовых машин

d) Распределенный сервис для коллекционирования, сбора, и перемещения больших массивов данных

3 Сколько уровней имеет лямбда-архитектура?

a) 2

b) 3
c) 4
d) 5
4 Что из этого является недостатком MapReduce?
a) Фиксированный алгоритм обработки данных
b) Масштабируемость
c) Отказоустойчивость
d) Возможность автоматического распараллеливания
5 Что из этого не является реализацией Hadoop?
a) Google MapReduce
b) Phoenix
c) GreenMint
d) Qizmt
6 Какая цель у NameNode в HDFS?
a) Хранить индекс того, какая часть данных находится в каком узле
b) Хранить имя файла, хранящегося в конкретном узле
c) Хранить индекс узла в котором хранится имя файла
d) Хранить имена узлов
7 Какие компоненты являются частями HDFS?
a) Task Tracker
b) Name Node и Data Node
c) Job Tracker и Task Tracker
d) Job Tracker, Task Tracker, Name Node и Data Node
8 Что такое GraphX?
a) Это библиотека для манипуляций над графами и выполнения с ними параллельных операций
b) Это базовый движок для крупномасштабной параллельной и распределенной обработки данных
c) Это проект Apache, который позиционируется как инструмент для «молниеносных кластерных вычислений»
d) Нереляционная распределенная база данных, исполняемая на верху HDFS
9 Какие из перечисленных пунктов являются достоинствами MapReduce?
a) Оптимальная производительность
b) Эффективное применение в маленьких кластерах с небольшим объемом данных
c) Масштабируемость
d) Отказоустойчивость
10 Какой из перечисленных узлов управляет другими?
a) Name Node
b) Data Node
c) Slave Node
d) Нет верного ответа
Раздел 4 Масштабирование и многоуровневое хранение данных

1 Разделение одной таблицы на разные сервера – это

a) Горизонтальное масштабирование

b) Вертикальное масштабированиие

c) Горизонтальный шардинг

d) Вертикальный шардинг

2 Выделение таблицы или группы таблиц на отдельный сервер это

a) Горизонтальное масштабирование

b) Вертикальное масштабированиие

c) Горизонтальный шардинг

d) Вертикальный шардинг

3 Выберите два определения, относящиеся к вертикальному масштабированию.

a) Возможность добавлять к системе новые узлы, серверы, процессоры для увеличения общей производительности

b) Возможность заменять в существующей вычислительной системе компоненты более мощными и быстрыми по мере роста требований и развития технологий

c) Разбиение системы на более мелкие структурные компоненты и разнесение их по отдельным физическим машинам (или их группам), и (или) увеличение количества серверов, параллельно выполняющих одну и ту же функцию

d) Увеличение производительности каждого компонента системы с целью повышения общей производительности

4 Что лежит в методологические основе идеи NoSQL?

a) Реляционная модель данных

b) Открытый исходный код

c) Хорошая вертикальная масштабируемость

d) Все ответы верны

5 Выберите одно неверное утверждение:

a) NoSQL базы данных удобны при хранении плохо структурированной информации

b) NoSQL базы данных разработаны для хранения структурированной информации

c) NoSQL базы данных эффективны для аналитической обработки данных

d) NoSQL базы данных не поддерживают join операции

6 Какой способ хранения данных используется в MongoDB?

a) Хэш-таблицы

b) Memory-mapped файлы

c) B+деревья

d) Неупорядоченные массивы

7 Основная идея NoSQL БД

a) Достичь линейной масштабируемости при большом числе узлов

b) Достичь большей скорости работы

c) Увеличение вместимости БД

d) Облегчение работы с БД

8 В основе идеи NoSQL лежит

a) Нереляционная модель данных

b) Хорошая вертикальная масштабируемость

c) Согласованность данных

d) Закрытый исходный код

9 Структура данных в таких базах представляет собой набор узлов, связанных между собой ссылками. При этом и узлы и ссылки могут обладать некоторым количеством атрибутов.

a) Wide Column Store

b) Document Store

c) Graph Store

d) ey/Value DB

10 Что означает термин NoSQL?

a) Не SQL

b) Не только SQL

c) Без SQL

d) SQL – плохо

Раздел 5 Визуализация данных и результатов анализа

1 Отображение зависимости значений одной величины от другой - это..

a) Матрица

b) График

c) Диаграмма

d) Карта

2 Что такое дедупликация данных?

a) Технология, при помощи которой обнаруживаются и исключаются избыточные данные в дисковом хранилище

b) Платформа для построения приложений, способных обрабатывать огромные объемы данных

c) Процесс, ориентированный на бизнес-пользователя и включающий доступ и исследование информации, ее анализ, выработку интуиции и понимания, которые ведут к улучшенному и неформальному принятию решений

d) Процесс создания (или выбора) модели для предсказания вероятности наступления некоторого события

3 Что из перечисленного помогает следить за эволюцией документа, над созданием которого работает одновременно большое количество авторов?

a) Пространственный поток

b) Исторический поток

c) Визуальный поток

d) Интерактивный поток

4 Как называется один из самых популярных языков сценариев?

a) G

b) B
c) R
d) P
5 Какое слово не входит в расшифровку Amazon S3?
a) Storage
b) Simple
c) Service
d) Scale
6 Какой тип лицензии у языка R?
a) Adware
b) Commercial CC
c) Open source
d) Shareware
7 Каков максимальный вес объекта, который можно загрузить через один запрос PUT в Amazon S3?
a) 64 Мб
b) 2 Гб
c) 5 Гб
d) 10 Тб
8 Как расшифровывается английская аббревиатура EDA?
a) Electronic design automation
b) European defence agency
c) Exploratory data analysis
d) Electronic design alliance
9 Достоинства Amazon S3:
a) Будет работать всегда
b) Нужно самостоятельно решать сложные задачи распределения файлов между серверами
c) Внезапные всплески популярности не приведут к отказу железа
d) Все вышеперечисленное
10 Что из перечисленного является быстрым прототипированием?
a) Создание нескольких различных визуальных представлений разнородных данных
b) Создание большого количества различных визуальных представлений одних и тех же данных
c) Создание большого количества различных визуальных представлений разнородных данных
d) Создание большого количества различных визуальных представлений противоречивых данных
Раздел 6 Сложные методы аналитики
1 Ошибкой первого рода в статистике является
a) Ошибка не обнаружить различия или связи, которые на самом деле существуют
b) Ошибка обнаружить различия или связи, которые на самом деле не существуют

c) Ошибка обнаружить различия или связи, которые на самом деле существуют

d) Нет верного ответа

2 Генеральная совокупность - это...

a) Полный набор объектов, с которым связана изучаемая проблема

b) Набор ассоциаций данной системы

c) Динамическое множество входных параметров

d) Окончательная совокупность объектов системы в целом

3 Чем отличаются ошибки первого и второго рода при принятии решений?

a) Ошибка первого рода значительнее, нежели второго

b) Ошибка второго рода не обнаруживает различия, которые есть, а первого обнаруживает, которых нет

c) Ошибка второго рода значительнее, нежели первого

d) Ошибка первого рода не обнаруживает различия, которых нет, а второго обнаруживает

4 Когда элементов слишком много, из множества отбирают выборку объемом N, исследуют ее, а затем изучаемые свойства выборки обобщаются на всю генеральную совокупность, такое обобщение называется:

a) Предварительная оценка

b) Общий вывод

c) Статистический вывод

d) Промежуточная оценка

5 Определённое предположение о распределении вероятностей, лежащем в основе наблюдаемой выборки данных, - это

a) Статистический критерий

b) Статистическая выборка

c) Статистическая гипотеза

d) Задача кластеризации

6 Графичекая характеристика качества бинарного классификатора ROC–кривая показывает зависимость

a) Величины TPR (доля верных положительных классификаций) от величины FPR(доля ложных положительных классификаций)

b) Величины FPR(доля ложных положительных классификаций)

от величины TPR (доля верных положительных классификаций)

c) Величины TNR (доля верных отрицательных классификаций) от величины FPR(доля ложных положительных классификаций)

d) Величины FPR(доля ложных положительных классификаций) от величины TNR (доля верных отрицательных классификаций)

7 Технология машинного обучения, когда нет ответов и требуется искать зависимости между объектами, называется …

a) Самостоятельное обучение

b) Обучение без учителя

c) Обучение с учителем

d) Обучение по зависимостям

8 К каким алгоритмам классификации относится метод ближайших соседей?

a) Метрическим

b) Логическим

c) Линейным

d) Нет верного ответа

9 Ошибкой второго рода в статистике является

a) Ошибка не обнаружить различия или связи, которые на самом деле существуют

b) Ошибка обнаружить различия или связи, которые на самом деле не существуют

c) Ошибка обнаружить различия или связи, которые на самом деле существуют

d) Нет верного ответа

10 Какая из перечисленных задач относится к задачам машинного обучения без учителя?

a) Задача классификации

b) Задача регрессии

c) Задача кластеризации

d) Нет верного ответа

Раздел 7 Анализ текста

1 Какие функции выполняет индексатор в поисковых сервисах?

a) Обработка текста

b) Поиск по индексу

c) Обработка поискового запроса

d) Сохранение в индексе

2 Из каких компонентов обычно состоит сервис поиска по тексту?

a) Анализатор

b) Индексатор

c) Поисковик

d) Нормализатор

3 Основной принцип технологий полнотекстового поиска.

a) Разделение входного текста на равные части для последовательного поиска в каждой.

b) Построение индекса для быстрого поиска соответствия по ключевым словам.

c) Сокращение объема входного текста путем удаления малозначащих слов.

d) Нет верного ответа.

4 Solr - одно из решений для полнотекстового поиска на базе

a) Elasticsearch

b) Lucene

c) R

d) MongoDB

5 Какие функции выполняет поисковик в поисковых сервисах?

a) Сохранение в индексе

b) Обработка текста

c) Поиск по индексу

d) Обработка поискового запроса

6 Что из перечисленного позволяет Word2Vec?

a) Получить векторные представления слов

b) Построить график распределения слов в корпусе текста

c) Сократить объем текста без потери исходного смысла

d) Определить связи между контекстами слов

7 Какие из перечисленных недостатков характерны для Lucene?

a) Отсутствие поддержки стемминга

b) Отсутствие API кроме Java

c) Невозможность объединения результатов при параллельном поиске

d) Низкая скорость индексации

8 Какой обработчик не входит в состав анализатора текста?

a) нормализации

b) токенизации

c) фильтрации полученных токенов

d) символьной фильтрации

9 Какой из перечисленных механизмов лучшим образом подходит для использования в составе сторонних продуктов?

a) Solr

b) R

c) Lucene

d) Elasticsearch

10 Что из перечисленного не является механизмом для реализации полнотекстового поиска?

a) Lucene

b) R

c) Elasticsearch

d) Solr

3 Выполнение практических работ

Задания для практических работ, теоретический и методический материал для самостоятельного изучения представлены на платформе «Открытое образование» в онлайн-курсе Наука о данных и аналитика больших объемов данных. – (https://openedu.ru/course/spbstu/BIGDATA/).

Раздел 1 Введение в большие данные

Необходимо выбрать предметную область, в рамках которой в дальнейшем будет проводиться работа. Для выбранной предметной области требуется сформулировать от 5 до 20 задач для проведения анализа. Задачи могут быть отнесены к следующим областям анализа:

анализ социальных сетей (Social Mining), анализ Интернет-ресурсов (Web Mining), анализ текста (Text Mining), анализ данных (Data Mining).

Раздел 2 Жизненный цикл аналитики данных

Определиться с источником данных и осуществить извлечение информации. Поскольку не предусмотрено предоставление готового набора данных для решения сформулированных задач анализа на предыдущем этапе, требуется самостоятельно сформировать собственный набор данных.

Раздел 3 Высокопроизводительные вычисления

Проработать архитектуру проектируемой системы. Проект подразумевает создание программного решения, позволяющего автоматически или полуавтоматически решать сформулированные ранее задачи анализа. В основе решения может быть заложена относительно простая, но функциональная и расширяемая модульная схема.

Раздел 4 Масштабирование и многоуровневое хранение данных

Определиться с подходом к хранению сформированного набора данных или набора данных, который прошел предварительную очистку и готов поступить на обработку. В проекте предполагается использование одной из распределенных баз данных, относящихся к типу NoSQL.

Раздел 5 Визуализация данных и результатов анализа

Начать реализацию модуля обработки данных. Данный модуль 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.

Раздел 6 Сложные методы аналитики

Продолжение работы над модулем обработки данных. Кроме Apache Hadoop для реализации модуля обработки данных может использоваться другая технология - Apache Spark.

Раздел 7 Анализ текста

Представить результаты проведенных исследований над данными в виде набора графиков и диаграмм, наглядно изображающих полученные выводы.

4 Оценивание обучающегося на экзамене

Промежуточная аттестация производится на основе текущего контроля успеваемости при условии выполнения всех заданий в семестре, предусмотренных фондом оценочных средств по дисциплине, и выполнения итогового контрольного теста с прокторингом на платформе «Открытое образование» в онлайн-курсе Наука о данных и аналитика больших объемов данных. – (https://openedu.ru/course/spbstu/BIGDATA/).