МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«ДАГЕСТАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Факультет математики и компьютерных наук

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Извлечение и анализ интернет данных

Кафедра прикладной математики
факультета математики и компьютерных наук

Образовательная программа бакалавриата:
01.03.05 - Статистика

Направленность (профиль) программы:
Анализ больших данных

Форма обучения:
очная

Статус дисциплины:
входит в часть ОПОП, формируемую участниками образовательных отношений;
дисциплина по выбору

Махачкала, 2023

Рабочая программа дисциплины «Извлечение и анализ интернет данных» составлена в 2023 году в соответствии с требованиями ФГОС ВО - бакалавриат по *направлению подготовки 01.03.05 Статистика* от 14.08.2020 г. №1032

Разработчик: кафедра прикладной математики:
Лугуева А.С, к.ф-м.н., доцент,

Рабочая программа дисциплины одобрена:
на заседании кафедры прикладной математики от «20» января 2023 г., протокол № 5
Зав. кафедрой _____ Кадиев Р.М.

на заседании Методической комиссии факультета математики и компьютерных наук от «25» января 2023 г., протокол № 4 .

Председатель _____ Ризаев М.К.

Рабочая программа дисциплины согласована с учебно-методическим управлением «___» _____2023 г.

Начальник УМУ _____ Гасангаджиева А.Г.
(подпись)

Аннотация рабочей программы дисциплины

Дисциплина «Извлечение и анализ интернет данных» является дисциплиной по выбору *бакалавриата* по направлению подготовки 01.03.05 - Статистика. Дисциплина реализуется на факультете математики и компьютерных наук ДГУ кафедрой прикладной математики.

Содержание дисциплины охватывает круг вопросов, связанных с обучением передовым методам, моделям, средствам и технологиям поиска и компьютерной обработки информации

Дисциплина нацелена на формирование следующих компетенций выпускника:

Профессиональных

- ПК - 1 Способен собирать, обрабатывать и интерпретировать данные современных научных исследований, необходимые для формирования выводов по соответствующим научным исследованиям;

Преподавание дисциплины предусматривает проведение следующих видов учебных занятий: *лекции, практические занятия, лабораторные занятия самостоятельная работа.*

Рабочая программа дисциплины предусматривает проведение следующих видов контроля успеваемости в форме *контрольной работы* и промежуточный контроль в форме *зачета*.

Объем дисциплины: 3 зачетные единицы, в том числе в академических часах по видам учебных занятий:

Очная форма обучения

Семестр	Учебные занятия в том числе:							СРС, в том числе экзамен	Форма промежуточной аттестации (зачет, дифференцированный зачет, экзамен)
	всего	Контактная работа обучающихся с преподавателем							
		всего	из них						
			Лекции	Лабораторные занятия	Практические занятия	КСР	консультации		
6	108	48	***16***	***16***	***16***			*60*	***Зачет***

1. Цели освоения дисциплины:

Целью дисциплины «Извлечение и анализ интернет данных» является – Целями освоения дисциплины Извлечение и анализ интернет-данных являются:

– Ознакомление студентов с основными способами извлечения информации из интернета и эффективного анализа этой информации

– Формирование у студентов практических навыков анализа и извлечения данных иработы с ними

Задачи дисциплины -дать знания о:

- истории и тенденциях развития информационно-поисковых систем, крупныхученых, участвовавших в их разработке,
- основных принципах обмена данными в глобальной сети Интернет;
- основных методах функционирования информационно-поисковых систем;
- основных современных инструментальных средствах их разработки;
- основных методах программирования поиска, как на стороне сервера, так и настороне клиента.

2.Место дисциплины в структуре ОПОП бакалавриата

Дисциплина «Извлечение и анализ интернет данных» входит входит в часть ОПОП *бакалавриата* по направлению подготовки **01.03.05 - Статистика**, формируемую участниками

образовательных отношений, дисциплина по выбору**.** Дисциплина реализуется на факультете математики и компьютерных наук ДГУ кафедрой прикладной математики.

При изучении дисциплины «Извлечение и анализ интернет данных» предполагается, что студент владеет основами статитики, математики, основами работы с большими данными.

Знания, навыки и умения, полученные студентами при изучении данной дисциплины, должны быть использованы в процессе изучения последующих дисциплин по учебному плану, связанных с реализацией цифровых компетенций.

Освоение дисциплины способствует формированию профессиональных компетенций и взаимодействуют с другими дисциплинами цикла.

3. Компетенции обучающегося, формируемые в результате освоения дисциплины (перечень планируемых результатов обучения).

Код и наименование компетенции из ОПОП	**Код и наименование индикатора достижения компетенций** *(в соответствии с ОПОП*	**Планируемые результаты обучения**	**Процедура освоения**
ПК-1. Способен собирать, обрабатывать и интерпретировать данные современных научных исследований, необходимые для формирования выводов по соответствующим научным исследованиям	ПК-1.1. Знает методы сбора и обработки данных, полученными в области математических и естественных наук, программирования и информационных технологий для формирования выводов по соответствующим научным исследованиям	Знает: стандартные методы и технические средства для статистических наблюдений. Умеет: применить стандартные методы и технические средства при статистических наблюдениях. Владеет: методами и техническими средствами для статистических наблюдений.	устный опрос, тестирование, письменный опрос
	ПК-1.2. Умеет собирать и обрабатывать данные, полученные в области математических и естественных наук, в области программирования и информационных технологий для формирования выводов по соответствующим научным исследованиям.	Знает: как собирать данные об объекте исследования и выбрать соответствующий инструментарий для обработки информации. Умеет: собирать исходные данные об объекте исследования и выбрать соответствующий инструментарий для обработки информации. Владеет: методами сбора данных об объекте исследования и выбора соответствующий инструментарий для обработки информации.	устный опрос, тестирование, письменный опрос
	ПК-1.3. Владеет навыками сбора и обработки данных, полученными в области математических и (или) естественных наук,	Знает: статистические методы обработки информации, в том числе с применением информационно-коммуникационных технологий.	устный опрос, тестирование, письменный опрос

	программирования и информационных технологий для формирования выводов по соответствующим научным исследованиям.	Умеет: применять статистические меты для обработки информации, в том числе с применением информационно-коммуникационных технологий. Владеет: статистическими методами обработки информации, в том числе с применением информационно-коммуникационных технологий	

4. Объем, структура и содержание дисциплины.

4.1. Объем дисциплины составляет 3 зачетные единицы, 108 академических часа.

4.2. Структура дисциплины.

4.2.1. Структура дисциплины в очной форме

№ п/п	Разделы и темы дисциплины	Семестр	Неделя семестра	Виды учебной работы, включая самостоятельную работу студентов и трудоемкость (в часах)					Формы текущего контроля успеваемости *(по неделям семестра)* Форма промежуточной аттестации *(по семестрам)*
				Лекции	Практические занятия	Лабораторные занятия	Итоговый контроль	СРС, в том числе экзамен	
	Модуль 1 **Аналитика в сети Интернет**								**Формы текущего контроля:** устные опросы, тестирование, реферат, доклады, **Форма промежуточной аттестации:** письменная контрольная работа
1	**Тема 1.** Генезис сети Интернет. .	6		2		2		8	
2	**Тема 2.** Структура WEB, Deep WEB	6		2	2	2		6	
3	**Тема 3.** Системы управления контентом.	6		2	2	2		6	
	Итого по модулю 1:			6	4	6		20	36
	Модуль 2 Методология сбора данных из сетевых источников								
6.	**Тема 4.** Технологии извлечения знаний из WEB - WEB-mining.	6		2	2	2		6	**Формы текущего контроля:** устные опросы, тестирование, реферат, доклады,
7	**Тема 5.** Понятие *data scraping* или	6		2	2	2		6	

	«срезание данных с поверхности». Классификация способов извлечения информации из WEB-источников.								**Форма промежуточной аттестации:** письменная контрольная работа
8	**Тема 6.** Модели информационного поиска.	6		2	2			8	
	Итого по модулю 2:			**6**	**6**	**4**		**20**	**36**
	Модуль 3 Типы информационных систем. Устройство ипринцип работы поисковых систем.								
6.	**Тема 7.** Типология, структура и функция информационных систем. Системы переработкиинформации. Типы информационных систем. Уточнение структуры информационных систем.Информационные системы Интернета.	6		2	2	2		6	**Формы текущего контроля:** устные опросы, тестирование, реферат, доклады, **Форма промежуточной аттестации:** письменная контрольная работа
7	**Тема 8.** Устройство и принцип работы поисковых систем. Автоматическое индексирование. Семантический вэб. Искусственный интеллект. Разработка ИПТ. Отраслевой тезаурус.	6		2	2	2		6	
8	**Тема 9.** Способы хранения больших данных в WEB	6			2	2		8	
	Итого по модулю 3:			**4**	**6**	**6**		**20**	**36**
	ИТОГО:	*6*		***16***	***16***	***16***		***60***	***108***

4.3. Содержание дисциплины, структурированное по темам (разделам).

4.3.1. Содержание лекционных занятий по дисциплине

Модуль 1. Аналитика в сети Интернет.

Тема 1. Генезис сети Интернет.

История создания Сети. Развитие электрических и электронных средств связи. ARPANET. Всемирная паутина. Развитие интернет в XXI веке. Организационная структура Интернета. Схема адресации в сети Интернет. Модель BOW TIE. Понятия и различия WEB 2.0- WEB 4.0.

Тема 2. Структура WEB, Deep WEB.

Невидимый WEB, его возможности и характеристики. Инструменты и технологии работы в невидимом WEB.

Тема 3. Системы управления контентом.

Проблемы, возникающие при поддержании актуальности информации на сайте. Определение CMS. Краткое описание CMS. Динамический и статический сайты. Характеристика контента. Создание контента. Управление автоматизированными деловыми процессами. Распространение контента.Персонализация и глобализация контента. Критерии классификации систем управления контентом. Простая CMS. Шаблонная CMS. Профессиональная CMS. Универсальная CMS. Функциональные и технологические возможности систем управления контентом. Требования к системам управления контентом. Вопросы, решаемые при выборе системыуправления контентом.

Модуль 2 Возможности и ограничения качественных методов в научных исследованиях

Тема 4. Технологии извлечения знаний из WEB -WEB-mining.

Определение понятий WEB Mining и Data Mining? Отличия между ними. Задачи и этапы извлечения знаний из WEB.Направления WEB-mining: Извлечение Web-контента (Web Content Mining); Извлечение Web-структур (Web Structure Mining); Исследование использования Web-ресурсов (Web Usage Mining)

Тема5. Понятие *data scraping* или «срезание данныхс поверхности». Классификация способов извлечения информации из WEB-источников.

Понятие бизнес- аналитического решения. Анализ журнала посещаемости сайта. Заказные статистические исследования. Определение профиля сайта. Определение перечня сайтов, посещаемых вашей аудиторией. Определение целевой аудитории сайта. Типы посетителей сайтов. Модели поведения посетителей сайта. Пользователи Интернет магазинов.

Тема 6. Модели информационного поиска.

Булева модель, векторная модель, вероятностная модель, гибридная модель. Математические особенности обработки информации разными моделями. Сферы их применения.

Модуль 3.Типы информационныхсистем. Устройство и принцип работы поисковых систем

Тема 7. Типология, структура и функция информационных систем.

Системы переработки информации. Типы информационных систем. Уточнение структуры информационных систем. Информационные системы Интернета.

Тема 8. Устройство и принцип работы поисковых систем.

Понятие поисковой системы. Принципы работы поисковых систем, которые нужно учитывать при продвижении сайта. Виды поисковых роботов. Порядок индексации сайтов. Порядок поисковой выдачи. Принципы алгоритмов выдачи поисковой системы Яндекс и Google. Выбор ключевых слов для продвижения сайта. Типы запросов по частотности. Типы запросов по степени конверсии. Понятие семантического ядра. Создание семантического ядра. Выбор ключевых страниц сайта. Распределение семантического ядра. Анализ сайтов конкурентов. Расчет сложности продвижения сайта. Выбор основной стратегии поискового продвижения сайта.

Тема 9. Способы хранения больших данных в WEB

Требования к хранилищам данных, OLTP и OLAP системы. Нереляционные базы данных.

4.3.2. Содержание практических занятий по дисциплине

Модуль 1. Аналитика в сети Интернет.

Тема 1. Генезис сети Интернет.

1. История создания Сети.
2. Развитие электрических и электронныхсредств связи.
3. ARPANET.
4. Всемирная паутина. Развитие интернет в XXI веке.
5. Организационная структура Интернета.
6 . Схема адресации в сети Интернет.
7. Модель BOW TIE. Понятия и различия WEB 2.0- WEB 4.0.

Тема 2. Структура WEB, Deep WEB.

1. Невидимый WEB, его возможности и характеристики.
2. Инструменты и технологии работы в невидимом WEB.

Тема 3. Системы управления контентом.

1. Проблемы, возникающие при поддержании актуальности информации на сайте.
2. Определение CMS. Краткое описание CMS.
3. Динамический и статический сайты.
4. Характеристика контента. Создание контента.
5. Управление автоматизированными деловыми процессами.

6 Распространение контента.

7. Персонализация и глобализация контента. Критерии классификации систем управления контентом.
8. Простая CMS. Шаблонная CMS. Профессиональная CMS. Универсальная CMS.
9. Функциональные и технологические возможности систем управления контентом. Требования к системам управления контентом. Вопросы, решаемые при выборе системыуправления контентом.

Модуль 2 Возможности и ограничения качественных методов в научных исследованиях

Тема 4. Технологии извлечения знаний из WEB -WEB-mining.

1. Определение понятий WEB Mining и Data Mining.
2. Отличия между ними. Задачи и этапы извлечения знаний из WEB.
3. Направления WEB-mining: Извлечение Web-контента (Web Content Mining);
4. Извлечение Web-структур (Web Structure Mining);
5. Исследование использования Web-ресурсов (Web Usage Mining)

Тема5. Понятие *data scraping* или «срезание данныхс поверхности». Классификация способов извлечения информации из WEB-источников.

1. Понятие бизнес- аналитического решения.
2 . Анализ журнала посещаемости сайта.
3. Заказные статистические исследования.
4. Определение профиля сайта.
5. Определение перечня сайтов, посещаемых вашей аудиторией.
6. Определение целевой аудитории сайта. Типы посетителей сайтов.
7. Модели поведения посетителей сайта. Пользователи Интернет магазинов.

Тема 6. Модели информационного поиска.

1. Булева модель, векторная модель, вероятностная модель, гибридная модель.
2. Математические особенности обработки информации разными моделями.
3. Сферы их применения.

Модуль 3. Типы информационныхсистем. Устройство и принцип работы поисковых систем

Тема 7. Типология, структура и функция информационных систем.

1. Системы переработки информации.
2. Типы информационных систем. Уточнение структуры информационных систем.
3. Информационные системы Интернета.

Тема 8. Устройство и принцип работы поисковых систем.

1. Понятие поисковой системы. Принципы работы поисковых систем, которые нужно учитывать при продвижении сайта.
2. Виды поисковых роботов.
3. Порядок индексации сайтов. Порядок поисковой выдачи.
4. Принципы алгоритмов выдачи поисковой системы Яндекс и Google.

5. Выбор ключевых слов для продвижения сайта. Типы запросов по частотности. Типы запросов по степени конверсии.
6. Понятие семантического ядра. Создание семантического ядра. Выбор ключевых страниц сайта. Распределение семантического ядра.
7. Анализ сайтов конкурентов.
8. Расчет сложности продвижения сайта. Выбор основной стратегии поискового продвижения сайта.

Тема 9. Способы хранения больших данных в WEB

1. Требования к хранилищам данных, OLTP и OLAP системы.
2. Нереляционные базы данных.

4.3.3. *Содержание лабораторных занятий по дисциплине*

Модуль 1. Аналитика в сети Интернет.

Тема 1. Генезис сети Интернет.

Лаб. работа 1. Вводное занятие. Настройка необходимого ПО и средыразработки.

Тема 2. Структура WEB, Deep WEB.

Лаб. работа 2. Составление запросов по теме магистерской работы, выполнениепоиска в открытых и закрытых сетевых источниках, сравнение эффективности поиска с помощьюразличных инструментов.

Тема 3. Системы управления контентом.

Лаб. работа 3. Обсуждение преимуществ и недостатков различных CMS, особенностей разработки WEB-ресурсов с их помощью.

Модуль 2 Возможности и ограничения качественных методов в научных исследованиях

Тема 4. *Технологии извлечения знаний из WEB -WEB-mining.*

Лаб. работа 4. Рассмотреть возможность использования любого из приведенных либо найденных способов извлечения информации с web страниц.

Тема5. *Понятие data scraping или «срезание данныхс поверхности». Классификация способов извлечения информации из WEB-источников.*

Лаб. работа 5. Используя любой из приведенных либо найденных способов извлечения информации с web страниц, разработать программу по сбору информации методами Web-scrappinga и продемонстрировать результат ее работы.

Тема 6. *Модели информационного поиска.*

Лаб. работа 6. Продемонстрировать результат работы, разработанной программы по сбору информации методами Web-scrappinga

Модуль 3.Типы информационныхсистем. Устройство и принцип работы поисковых систем

Тема 7. *Типология, структура и функция информационных систем.*

Лаб. работа 7. Определение и анализ характеристик выбранной поисковой системы: Google, Yandex, Rambler

Тема 8. *Устройство и принцип работы поисковых систем.*

Лаб. работа 8. Определение и анализ характеристик выбранной поисковой системы: Yahoo, Bing, AltaVista.

Тема 9. *Способы хранения больших данных в WEB*

5. **Образовательные технологии**

Лекции проводятся с использованием меловой доски и мела. Параллельно материал транслируется на экран с помощью мультимедийного проектора. Для проведения лекционных занятий необходима аудитория, оснащенная мультимедиа-проектором, экраном, доской, ноутбуком (с программным обеспечение для демонстрации слайд-презентаций).

Для проведения практических занятий необходима аудитория на 25 человек, оснащена доской, компьютерами.

На лекционном и практическом занятиях посредством мультимедийных средств широко используется **демонстрационный материал**, который усиливает ощущения и восприятия обучаемого.

В частности, при изучении дисциплины предусмотрено применение следующих образовательных технологий:

– *Лекция-беседа*, являющаяся наиболее распространенной и сравнительно простой формой активного вовлечения студентов в учебный процесс. Эта лекция предполагает непосредственный контакт преподавателя с аудиторией. Преимущество лекции-беседы состоит в том, что она позволяет привлекать внимание студентов к наиболее важным. вопросам темы, определять содержание и темп изложения учебного материала с учетом особенностей студентов.

– *Проблемная лекция*, определяющим признаком которой является постановка и разрешение учебных проблем с различной степенью приобщения к этому слушателей. Такое занятие начинается с вопросов, с постановки проблемы, которую необходимо решить в ходе изложения материала.

– *Лекция-визуализация*, во время которой происходит переработка учебной информации по теме лекционного занятия в визуальную форму для представления студентам через технические средства обучения или вручную (схемы, рисунки, чертежи и т.п.).

Презентация – представление студентом наработанной информации по заданной тематике в виде набора слайдов и спецэффектов, подготовленных в выбранной программе.

– *Творческие задания* – самостоятельная творческая деятельность студента, в которой он реализует свой личностный потенциал, демонстрирует умение грамотно и ясно выражать свои мысли, идеи.

– *Компьютерные технологии* (компьютерный опрос, лекция – презентация, доклады студентов в сопровождении мультимедиа);

– *Диалоговые технологии* (опрос, взаимоопрос, дискуссия между студентами, дискуссия преподавателя и студентов);

– Технологии на основе метода *опережающего обучения* и др.

В ходе изучения дисциплины предусматриваются активные и интерактивные формы проведения занятий, в частности, с использованием разнообразных методов организации и осуществления:

– *учебно-познавательной деятельности* (словесные, наглядные и практические методы передачи информации, проблемные лекции и др.);
– *стимулирования и мотивации учебно-познавательной деятельности* (дискуссии, самостоятельные исследования по обозначенной проблематике, публикация статьи и др.);
– *контроля и самоконтроля* (индивидуального и фронтального, устного и письменного опроса, зачета).

6. Учебно-методическое обеспечение самостоятельной работы студентов.

Самостоятельная работа рассматривается как форма организации обучения, которая способна обеспечивать самостоятельный поиск необходимой информации, творческое восприятие и осмысление учебного материала в ходе аудиторных занятий, разнообразные формы познавательной деятельности студентов на занятиях и во внеаудиторное время, развитие аналитических способностей, навыков контроля и планирования учебного времени, выработку умений и навыков рациональной организации учебного труда. Она является формой организации образовательного процесса, стимулирующей активность, самостоятельность и познавательный интерес студентов, а также одним из обязательных видов образовательной деятельности, обеспечивающей реализацию требований Федеральных государственных стандартов высшего профессионального образования (ФГОС).

Самостоятельная работа студента выполняется по заданию и при методическом руководстве преподавателя и реализуется непосредственно в процессе аудиторных занятий – на лекциях и практических занятиях, а также вне аудитории – в библиотеке, на кафедре, дома и т.д.

Аудиторная самостоятельная работа студента осуществляется на лекционных и практических занятиях в форме выполнения различных заданий и научных работ. Внеаудиторная самостоятельная работа студента традиционно включает такие виды деятельности, как проработка ранее прослушанного лекционного материала, конспектирование программного материала по учебникам, подготовка доклада, выполнение реферата, поиск наглядного материала, выполнение предложенных преподавателем заданий в виртуальной обучающей системе в режиме on-line и т.д.

Самостоятельная работа студента должна быть ориентирована на поиск и анализ учебного и научного материалов для подготовки к работе на семинарском занятии и обсуждения заранее заданных и возникающих в ходе занятия вопросов.

Эффективность и конечный результат самостоятельной работы студента зависит от умения работать с научной и учебной литературой, источниками и информацией в сети Интернет по указанным адресам.

При изучении дисциплины «***Извлечение и анализ интернет данных***» используются следующие виды самостоятельной работы студентов:

При оценивании результатов освоения дисциплины (текущей и промежуточной аттестации) применяется балльно-рейтинговая система, внедренная в Дагестанском государственном университете. В качестве оценочных средств на протяжении семестра используется тестирование, контрольные работы студентов, творческая работа, итоговое испытание.

Основными видами самостоятельной работы студентов являются:

1. изучение рекомендованной литературы, поиск дополнительного материала;
2. работа над темами для самостоятельного изучения;
3. подготовка к ***зачету***.

Темы, виды и содержание самостоятельной работы по дисциплине Темы, виды и содержание самостоятельной работы по дисциплине

Темы	Виды и содержание самостоятельной работы	Форма контроля
Тема 1. Генезис сети Интернет. .	1. Проработка конспекта лекций. 2. Поиск и анализ дополнительной литературы.	Устный опрос, тестирование, презентация.
Тема 2. Структура WEB, Deep WEB	1. Проработка конспекта лекций, изучение учебной и научной литературы и интернет ресурсов; 2. Подготовка к слабораторному занятию по теме, составление конспекта.	Устный опрос, тестирование, презентация.
Тема 3. Системы управления контентом.	1. Проработка конспекта лекций, изучение учебной и научной литературы и интернет ресурсов; 2. Поиск и анализ дополнительной литературы.	Устный опрос, тестирование, презентация.
Тема 4. Технологии извлечения знаний из WEB -WEB-mining.	1. Проработка конспекта лекций, изучение учебной и научной литературы и интернет ресурсов; 2. Подготовить реферат по теме.	Устный опрос, тестирование, презентация..
Тема 5. Понятие *data scraping* или «срезание данныхс поверхности». Классификация способов извлечения информации из WEB-источников.	1. Проработка конспекта лекций. 2. Поиск и анализ дополнительной литературы.	Устный опрос, тестирование, презентация.
Тема 6. Модели информационного поиска.	1. Проработка конспекта лекций, изучение учебной и научной литературы и интернет ресурсов; 2. Разработать электронную презентацию	Устный опрос, тестирование, презентация..
Тема 7. Типология, структура и функция информационных систем. Системы переработкиинформации. Типы информационных систем.	1. Проработка конспекта лекций, изучение учебной и научной литературы и интернет ресурсов; 2. Подготовить реферат по теме.	Устный опрос, тестирование, презентация.

Уточнение структуры информационных систем.Информационные системы Интернета.		
Тема 8. Устройство и принцип работы поисковых систем. Автоматическое индексирование.Семантический вэб. Искусственный интеллект.Разработка ИПТ. Отраслевой тезаурус.	1. Проработка конспекта лекций. 2. Поиск и анализ дополнительной литературы.	Устный опрос, тестирование, презентация.
Тема 9. Способы хранения больших данных в WEB	1. Проработка конспекта лекций. 2. Поиск и анализ дополнительной литературы.	Устный опрос, тестирование, презентация.

7. Фонд оценочных средств для проведения текущего контроля успеваемости, промежуточной аттестации по итогам освоения дисциплины.

7.1. **Типовые контрольные задания**

Темы для подготовки презентаций

1. Основы анализа данных в Python
2. Визуализация данных в Python: библиотеки matplotlib, seaborn, plotly
3. Продвинутые инструменты для анализа данных
4. Парсинг открытых данных в различных форматах (xml/json/html)
5. Основы машинного обучения и практика применения
6. Извлечение данных сайта Вконтакте и изучение влияния социальных сетей на поведение в реальной жизни
7. Извлечение и анализданных Московской биржи

Примерный тест по дисциплине

Вопрос 1:
Показать правильные ответы
Непрерывные данные — это …
Варианты ответа:

а) данные, значения которых могут принимать какое угодно значение в некотором интервале

б) данные являющиеся значениями признака, общее число которых конечно либо бесконечно, но может быть подсчитано при помощи натуральных чисел от одного до бесконечности
в) числовые данные, упорядоченные по значению какого-либо признаками
г) логически взаимосвязанные между собой сведения, характеризующие определенный объект, процесс или явление

Вопрос 2:
Перевод Knowledge Discovery in Databases (KDD):
Варианты ответа:
а) извлечение данных из неструктурированных массивов
б) извлечение знаний из баз данных
в) тиражирование знаний
г) «раскопка» данных

Вопрос 3:
Статистический пакет можно отнести к классу аналитических платформ:
Варианты ответа:
а) Нет
б) Да

Вопрос 4:
Особенности данных, накапливаемых в компаниях:
Варианты ответа:
а) Как правило, данные содержат ошибки, аномалии и пропуски
б) Почти всегда носят неполный, фрагментарный характер
в) Данные редко накапливаются специально для решения задач анализа
г) Данные всегда представлены в структурированной форме
д) Нередко имеют большой объем

Вопрос 5:
Выберите неверный вариант:
Варианты ответа:
а) Эксперт выдвигает гипотезы и строит модели для проверки достоверности гипотез
б) Аналитик – это специалист в области анализа и моделирования
в) Эксперт является связующим звеном между специалистами разных уровней и областей
г) Эксперт – это специалист предметной области, профессионал, который за годы обучения и практической деятельности научился эффективно решать задачи, относящиеся к конкретной предметной области

Вопрос 6:
Числовые данные могут быть дискретного вида при решении задачи анализа:
Варианты ответа:
а) Да
б) Нет

Вопрос 7:
Подход моделирования, при котором отправной точкой являются данные, характеризующие исследуемый объект, и модель «подстраивается» под действительность – это … подход:
Варианты ответа:
а) аналитический

б) графический
в) интеллектуальный
г) информационный

Вопрос 8:
Самая распространенная модель хранения структурированных данных:
Варианты ответа:
а) текст
б) граф
в) таблица
г) дерево
д) матрица

Вопрос 9:
Перевод Data Mining:
Варианты ответа:
а) извлечение данных из неструктурированных массивов
б) тиражирование знаний
в) извлечение знаний из баз данных
г) «раскопка» данных

Вопрос 10:
Принципы, которым необходимо следовать при сборе данных:
Варианты ответа (один или несколько):
а) Собирать данные за два последних периода
б) Абстрагироваться от существующих информационных систем и имеющихся в наличии данных
в) Описать все факторы, возможно влияющие на анализируемый процесс/объект
г) Собирать только структурированные данные
д) Собрать все легкодоступные факторы
е) Собирать только слабоструктурированные данные
ё) Собирать только данные за последний год
ж) Экспертно оценить значимость каждого фактора
з) Обязательно собрать наиболее значимые с точки зрения экспертов факторы

Перечень вопросов для подготовки к зачету

1. Опишите структуру, пропорции, охарактеризуйте размеры и динамику WEB.
2. Понятие “Сильной связности» WEB-графа, типы его узлов. Какому функциональному закону подчиняются сети «тесного мира»?
3. Закономерности и ограничения модели Bow Tie.
4. Понятие WEB 2.0.
5. Deep WEB. Какие ресурсы его составляют. Какими средствами его можноисследовать.
6. Понятия Web Mining и Web Analytics. Этапы аналитики в соответствии состандартом CRISP-DM.
7. Задачи Data Mining. Направления Data Mining.
8. Понятие и задачи Web Content Mining.
9. Перечислите и охарактеризуйте средства WEB scraping.
10. Методы Text Mining в приложении к специфике WWW.
11. Методологии Web Graph Mining для подхода Web Structure Mining.

12. Основные задачи Web Usage Mining, средства их решения, назначение кластерного анализа в контексте Web Usage Mining.
13. Классификация способов извлечения информации из WEB-источников.
14. Задачи Web-scrapping, механизм его работы. Разновидность методов Web-scrapping.
15. Этапы работы поисковой системы. Компоненты поискового движка.
16. Как работают алгоритмы индексирования. Необходимость ранжирования и задачи машинного обучения в приложении к информационному поиску.
17. Охарактеризуйте модели информационного поиска.
18. Изложите подробно принцип булевой модели информационного поиска (ИП),возможные средства оптимизации запроса.
19. Суть векторной и вероятностной моделей ИП, их достоинства и недостатки.
20. Назовите и кратко охарактеризуйте этапы нормализации текста перед индексацией.
21. Перечислите и дайте краткую характеристику методов лингвистического анализа.
22. Способы хранения словарей. Способы нечеткого поиска.
23. Технология Map-Reduce, механизмы работы, примеры использования. Какобеспечивается отказоустойчивость Map-Reduce.
24. Технология Hadoop. MapReduce в Hadoop. Структура программы в Hadoop.
25. Хранилища Больших данных. Примеры распределенных хранилищ.
26. NoSQL, типы NoSQL баз данных. Теорема CAP.
27. Понятия OLAP и OLTP. Характеристики Больших данных.

7.2. Методические материалы, определяющие процедуру оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций.

1. Общий результат выводится как интегральная оценка, складывающая из текущего контроля – 50% и промежуточного контроля – 50%.
   Текущий контроль по дисциплине включает:

- посещение занятий– 10 баллов,
- участие на практических занятиях– 20 баллов,
- выполнение самостоятельных, контрольных работ– 20 баллов.

Промежуточный контроль по дисциплине включает:

- письменная контрольная работа - 50 баллов.

2. Критерии оценок при проведении текущего контроля успеваемости

- Выполнение контрольной работы**:**

оценка «отлично» - выставляется студенту, если студент дал подробные ответы на все заданные вопросы. При этом студент должен показать знания не только из основной литературы, но и знания из дополнительной литературы, сети Internet;

оценка «хорошо» - выставляется студенту, если студент дал полные ответы на все вопросы, показав знания иосновной литературы. При этом студент допустил несущественнее недочеты в ответах и незначительные нарушения логики изложения материала;

оценка «удовлетворительно»: знание и понимание основного материала, наличие несущественныхошибок (не более 50%) при неспособности их последовательного и логического изложения, вызывает затруднение использование терминологии дисциплины;

оценка «неудовлетворительно»: непонимание сущности вопросов, грубые существенные ошибки вответе, отсутствие способности к письменному изложению материала.

- Критерии оценки коллоквиума:

оценка «отлично»: ответ полный, правильный, самостоятельный; материал изложен в

определенной логической последовательности, демонстрируется многосторонность подходов, многоаспектность обсуждения проблемы, умение находить рациональные пути решения задач, устанавливать причинно- следственные связи, в логическом рассуждении при решении задачи, графических построениях нет ошибок, задача решена рациональным способом с корректным использованием необходимых величин, получен верный ответ. Верные ответы даны на 86-100%

оценка «хорошо»: дан полный, правильный ответ на основе изученных понятий, но допускаются несущественные ошибки. Верные ответы даны на 66-85%.

оценка «удовлетворительно»: дан полный ответ, но при этом есть существенные ошибки указывающиена неумение использовать теоретические знания и умения при решении поставленных задач. Данные пробелы в знаниях не препятствуют дальнейшему обучению. Верные ответы даны на 51-65%

оценка «неудовлетворительно»: ответ обнаруживает незнание основного (порогового) содержанияучебного материала Верные ответы даны менее 50%.

Контроль освоения дисциплины и оценка знаний обучающихся на **зачете** производится в соответствии с Положением о проведении текущего контроля успеваемости и промежуточной аттестации обучающихся ДГУ и его филиалов.

Оценка «зачтено» ставится в том случае, когда студент обнаруживает систематическоеи глубокое знание основного содержания программного материала по дисциплине «Введение в ИТ», умеет свободно ориентироваться в вопросе. Ответ полный. Выдвинутые положения аргументированы и иллюстрированы примерами. Материал изложен в определенной логической последовательности, осознанно, литературным языком, с использованием современных научных терминов. Студент уверенно отвечает на дополнительные вопросы;

оценка «незачтено»: ответ обнаруживает незнание основного (порогвого) содержания учебного материала. менее 50%, уровень не сформирован.

Шкала диапазона для перевода рейтингового балла по дисциплине с учётом итогового контроля:

0 – 50 баллов – «незачтено»;

51 – 100 баллов – «зачтено»;

8. Учебно-методическое обеспечение дисциплины.

а) адрес сайта курса:

1. Сайт кафедры прикладной математики ДГУ:
http://cathedra.dgu.ru/OfTheDepartment.aspx?id=7
2. Образовательный блог: https://chislen-met.blogspot.com/

б) Основная литература:

1. Синица С.Г. Веб-программирование и веб-сервисы – учебное пособие, КубГУ, 2013. (28 экз. в библиотеке КубГУ).
2. Щербаков, А. Интернет-аналитика: поиск и оценка информации в web-ресурсах : практическое пособие / А. Щербаков. – Москва : Книжный мир, 2012. – 78 с. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=89693 – ISBN 978-5-8041-0569-4. – Текст : электронный.
3. Жуковский, О.И. Информационные технологии и анализ данных : учебное пособие / О.И. Жуковский ; Министерство образования и науки Российской Федерации, Томский Государственный Университет Систем Управления и Радиоэлектроники (ТУСУР). - Томск : Эль Контент, 2014. - 130 с. : схем., ил. - Библиогр.: с. 126. [Электронный ресурс].- URL: http://biblioclub.ru/index.php?page=book&id=480500

***в)* Дополнительная литература:**

1. Эзрахи, А. Виртуальная конкуренция: посулы и опасности алгоритмической экономики : учебник / А. Эзрахи, М. Стаки ; пер. с англ. под науч. ред. А. Резвова ; Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации. – Москва : Дело, 2022. – 384 с. – (Академическая книга). – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=685894 – Библиогр. в кн. – ISBN 978-5-85006-341-2. – Текст : электронный.
2. Оверби, Х. Цифровая экономика: как информационно-коммуникационные технологии влияют на рынки, бизнес и инновации : учебник / Х. Оверби, Я. А. Одестад ; под науч. ред. М. И. Левина ; пер. с англ. И. М. Агеевой ; пер. на англ. Н. В. Шиловой ; Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации. – Москва : Дело, 2022. – 288 с. : ил. – (Академическая книга). – Режим доступа: по подписке. – URL: – Библиогр.: с. 239-244. – ISBN 978-5-85006-391-7. – Текст : электронный.

9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины.

1. Университетская библиотека online : [электронно-библиотечная система] / ООО «ДиректМедиа». — Москва, 2001 — . — URL: http://www.biblioclub.ru — Режим доступа: по подписке. — Текст: электронный
2. eLIBRARY.RU [Электронный ресурс]: электронная библиотека / Науч. электрон. б-ка. — Москва, 1999 – . Режим доступа: http://elibrary.ru/defaultx.– Яз. рус., англ.
3. Электронный каталог НБ ДГУ [Электронный ресурс]: база данных содержит сведения о всех видах лит, поступающих в фонд НБ ДГУ/Дагестанский гос. ун-т. – Махачкала, 2010 – Режим доступа: http://elib.dgu.ru, свободный
4. КонсультантПлюс — студенту и преподавателю : [справочно-правовая система] / ООО Компания «КонсультантПлюс». — Москва, 1997 — . — URL: https://student.consultant.ru/card/ — Режим доступа: для зарегистрир. пользователей. — Текст : электронный
5. Book.ru : электронно-библиотечная система / ООО «КноРус Медиа». — Москва, 2010 — . — URL: https://www.book.ru/ — Режим доступа: по подписке. — Текст: электронный.

10. Методические указания для обучающихся по освоению дисциплины.

Перечень учебно-методических изданий, рекомендуемых студентам, для подготовки к занятиям представлен в разделе «Учебно-методическое обеспечение. Литература».

Для успешного освоения курса студентам рекомендуется проводить самостоятельный разбор материалов семинарских занятий в течении семестра. В случае затруднений в понимании и освоении каких-либо тем решать дополнительные задания из учебных пособий, рекомендуемых к данному курсу.

Важнейшей задачей учебного процесса в университете является формирование у студента общекультурных и профессиональных компетенций, в том числе способностей к саморазвитию и самообразованию, а также умений творчески мыслить и принимать решения на должном уровне. Выработка этих компетенций возможна только при условии активной учебно-познавательной деятельности самого студента на всём протяжении образовательного процесса с использованием интерактивных технологий.

Такие виды учебно-познавательной деятельности студента как лекции, семинарские занятия и самостоятельная работа составляют систему вузовского образования.

Лекция является главным звеном дидактического цикла обучения в отечественной высшей школе. Несмотря на развитие современных технологий и появление новых методик обучения лекция остаётся основной формой учебного процесса. Она представляет собой последовательное и систематическое изложение учебного материала, разбор какой-либо узловой проблемы. Вузовская лекция ориентирована на формирование у студентов информативной основы для последующего глубокого усвоения материала методом самостоятельной работы, призвана помочь студенту сформировать собственный взгляд на ту или иную проблему.

При изучении дисциплины рекомендуется рейтинговая технология обучения, которая позволяет реализовать комплексную систему оценивания учебных достижений студентов. Текущие оценки усредняются на протяжении семестра при изучении модулей. Комплексность означает учет всех форм учебной и творческой работы студента в течение семестра.

Рейтинг направлен на повышение ритмичности и эффективности самостоятельной работы студентов. Он основывается на широком использовании тестов и заинтересованности каждого студента в получении более высокой оценки знаний по дисциплине.

Рейтинговый балл студента на каждом занятии зависит от его инициативности, качества выполненной работы, аргументированности выступления, характера использованного материала и т.д. Уровень усвоения материала напрямую зависит от внеаудиторной самостоятельной работы, которая традиционно такие формы деятельности, как выполнение письменного домашнего задания, подготовка к разбору ранее прослушанного лекционного материала, подготовка доклада и выполнение реферата.

11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень программного обеспечения и информационных справочных систем.

Информационные средства обучения: электронные учебники, презентации, технические средства предъявления информации (многофункциональный мультимедийный комплекс) и контроля знаний (тестовые системы). Электронные ресурсы Научной библиотеки ДГУ. Электронно-образовательные ресурсы Дагестанского государственного университета.

Для успешного освоения дисциплины, обучающийся использует следующие программные средства: WINDOWSXP, пакет MSOFFICE.

12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине.

Компьютерный класс, аудитория для проведения лекционных и практических занятий и самостоятельной работы средствами оборудованная оргтехникой, персональными компьютерами, объединенными в сеть с выходом в Интернет; установленное лицензионное и свободное программное обеспечение