МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Кузбасский гуманитарно-педагогический институт
федерального государственного бюджетного образовательного учреждения
высшего образования
«Кемеровский государственный университет»

Факультет информатики, математики и экономики

УТВЕРЖДАЮ
Декан А.В. Фомина
«09» февраля 2023 г.

Рабочая программа дисциплины

К.М.07.04 Технологии работы с открытыми данными

Направление подготовки
01.03.02 Прикладная математика и информатика

Направленность (профиль) подготовки
ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

Программа бакалавриата

Квалификация выпускника
бакалавр

Форма обучения
Очная

Год набора 2022

Новокузнецк 2023

Оглавление

1 Цель дисциплины.

В результате освоения данной дисциплины у обучающегося должны быть сформированы компетенции основной профессиональной образовательной программы бакалавриата (далее - ОПОП): ПК-1.

Содержание компетенций как планируемых результатов обучения по дисциплине см. таблицы 1 и 2.

1.1 Формируемые компетенции

Таблица 1 - Формируемые дисциплиной компетенции

Наименование вида компетенции *(универсальная, общепрофессиональная, профессиональная)*	Наименование категории (группы) компетенций	Код и название компетенции
профессиональная		ПК-1 Способен проводить аналитические исследования с применением технологий больших данных

1.2 Индикаторы достижения компетенций

Таблица 2 – Индикаторы достижения компетенций, формируемые дисциплиной

Код и название компетенции	Индикаторы достижения компетенции по ОПОП	Дисциплины и практики, формирующие компетенцию ОПОП
ПК-1 Способен проводить аналитические исследования с применением технологий больших данных	ПК 1.1 Способен осуществлять выявление, формирование и согласование требований к результатам аналитических работ с применением технологий больших данных ПК 1.2 Способен планировать и организовывать аналитические работы с использованием технологий больших данных ПК 1.3 Способен подготавливать данные для проведения аналитических работ по исследованию больших данных ПК 1.4 Способен проводить аналитические исследования с применением технологий больших данных в соответствии с требованиями заказчика	К.М.07.01 Машинное обучение К.М.07.02 Аналитика данных К.М.07.03 Технологии работы с большими данными **К.М.07.04 Технологии работы с открытыми данными** К.М.07.ДВ.01.01 Машинное обучение с подкреплением/ К.М.07.ДВ.01.02 Глубокое обучение К.М.09.06(Пд) Преддипломная практика

1.3 Знания, умения, навыки (ЗУВ) по дисциплине

Таблица 3 – Знания, умения, навыки, формируемые дисциплиной

Код и название компетенции	Индикаторы достижения компетенции, закрепленные за дисциплиной	Знания, умения, навыки (ЗУВ), формируемые дисциплиной

Код и название компетенции	Индикаторы достижения компетенции, закрепленные за дисциплиной	Знания, умения, навыки (ЗУВ), формируемые дисциплиной
ПК-1 Способен проводить аналитические исследования с применением технологий больших данных	ПК 1.3 Способен подготавливать данные для проведения аналитических работ по исследованию больших данных	**Знать:** - методы эффективного поиска больших данных в открытых сетевых источниках; - приемы разработки информационно-поисковых систем для нахождения данных на стороне сервера или клиента. **Уметь:** - производить эффективный поиск больших данных, в том числе, в областях знаний, непосредственно не связанных со сферой деятельности; - подготавливать данные, полученные из открытых сетевых источников, для проведения аналитических работ по исследованию больших данных. **Владеть:** - навыками эффективного поиска больших данных.

2 Объём и трудоёмкость дисциплины по видам учебных занятий. Формы промежуточной аттестации.

Таблица 4 – Объем и трудоемкость дисциплины по видам учебных занятий

Общая трудоемкость и виды учебной работы по дисциплине, проводимые в разных формах	Объём часов по формам обучения
	ОФО
1 Общая трудоемкость дисциплины	72
2 Контактная работа обучающихся с преподавателем (по видам учебных занятий) (всего)	42
Аудиторная работа (всего):	42
в том числе:	
лекции	6
лабораторные работы	36
Внеаудиторная работа (всего):	
3 Самостоятельная работа обучающихся (всего)	30
4 Промежуточная аттестация обучающегося – зачет (8 семестр)	

3. Учебно-тематический план и содержание дисциплины.

3.1 Учебно-тематический план

Таблица 5 - Учебно-тематический план очной формы обучения

№ недели п/п	Разделы и темы дисциплины по занятиям	Общая трудоёмкость *(всего час.)*	Трудоемкость занятий (час.)			Формы текущего контроля и промежуточной аттестации успеваемости
			ОФО			
			Аудиторн. занятия		СРС	
			лекц.	лаб.		

№ недели п/п	Разделы и темы дисциплины по занятиям	Общая трудоёмкость *(всего час.)*	Трудоемкость занятий (час.)			Формы текущего контроля и промежуточной аттестации успеваемости
			ОФО			
			Аудиторн. занятия		СРС	
			лекц.	лаб.		
Семестр 8						
1	Аналитика в сети Интернет	24	2	12	10	Лабораторные работы
2	Устройство и принцип работы поисковых систем	24	2	12	10	Лабораторные работы
3	Методология сбора и анализа данных из сетевых источников	24	2	12	10	Лабораторные работы Тест
	Промежуточная аттестация – *зачет*					
ИТОГО по семестру 8		**72**	**6**	**36**	**30**	

3.2. Содержание занятий по видам учебной работы

Таблица 6 – Содержание дисциплины

№ п/п	Наименование раздела, темы дисциплины	Содержание занятия
	Семестр 8	
Содержание лекционного курса		
1	Аналитика в сети Интернет	ARPANET. Всемирная паутина. Развитие интернета в XXI веке. Организационная структура интернета. Схема адресации в сети интернет. Модель BOW TIE. Понятия и различия WEB 2.0- WEB 4.0. Невидимый WEB, его возможности и характеристики. Инструменты и технологии работы в невидимом WEB. 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 системы управления контентом
2	Устройство и принцип работы поисковых систем	Типология, структура и функция информационных систем. Системы переработки информации. Типы информационных систем. Уточнение структуры информационных систем. Информационные системы Интернета. 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.

№ п/п	Наименование раздела, темы дисциплины	Содержание занятия
		Создание семантического ядра. Выбор ключевых страниц сайта. Распределение семантического ядра. Анализ сайтов конкурентов. Расчет сложности продвижения сайта. Выбор основной стратегии поискового продвижения сайта. Способы хранения данных в WEB.
3	Методология сбора и анализа данных из сетевых источников	Технологии извлечения знаний из WEB - WEB-mining. Определение понятия WEB Mining 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) Понятие data scraping или «срезание данных с поверхности». Понятие бизнесаналитического 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. Модели информационного поиска. Булева модель, векторная модель, вероятностная модель, гибридная модель. Математические особенности обработки информации разными моделями. Сферы их применения
Содержание лабораторных занятий		
1	Аналитика в сети Интернет	Выполнение поиска в открытых и закрытых сетевых источниках, сравнение эффективности поиска с помощью различных инструментов. Системы управления контентом. Сравнение преимуществ и недостатков различных CMS, особенностей разработки WEB-ресурсов с их помощью.
2	Устройство и принцип работы поисковых систем	Определение и анализ характеристик поисковых систем: Google, Yandex, Rambler, Yahoo, Bing, AltaVista. Устройство хранения данных в WEB.
3	Методология сбора и анализа данных из сетевых источников	Технологии извлечения знаний из WEB – WEB Mining. Понятие data scraping или «срезание данных с поверхности».
	Промежуточная аттестация - экзамен	

4 Порядок оценивания успеваемости и сформированности компетенций обучающегося в текущей и промежуточной аттестации.

Для положительной оценки по результатам освоения дисциплины обучающемуся необходимо выполнить все установленные виды учебной работы. Оценка результатов работы обучающегося в баллах (по видам) приведена в таблице 7.

Таблица 7 - Балльно-рейтинговая оценка результатов учебной работы обучающихся по видам (БРС)

Учебная работа (виды)	Сумма баллов	Виды и результаты учебной работы	Оценка в аттестации	Баллы (17 недель)
Текущая учебная работа в семестре (Посещение занятий по расписанию и выполнение заданий)	**60**	Лабораторные работы (отчет о выполнении лабораторной работы) (18 работ).	**1,5 балла** – выполнение работы на 51-65% **3 балла** – выполнение работы на 85,1-100%	27 - 54
		Тест	**4 балла** – выполнение работы на 51-65% **6 баллов** – выполнение работы на 85,1-100%	4-6
Итого по текущей работе в семестре				31 - 60
Промежуточная аттестация (экзамен)	40	Ответ на вопрос	**5 баллов** (пороговое значение) **8 баллов** (максимальное значение)	5-8
		Ответ на вопрос	**5 баллов** (пороговое значение) **8 баллов** (максимальное значение)	5-8
		Решение задачи	**10 баллов** (пороговое значение) **24 балла** (максимальное значение)	10 - 24
Итого по промежуточной аттестации (экзамену)				20 – 40 б.
Суммарная оценка по дисциплине: Сумма баллов текущей и промежуточной аттестации				51 – 100 б.

В промежуточной аттестации оценка выставляется в ведомость в 100-балльной шкале и в буквенном эквиваленте (таблица 8)

Таблица 8 – Соотнесение 100-балльной шкалы и буквенного эквивалента оценки

Сумма набранных баллов	Уровни освоения дисциплины и компетенций	*Экзамен*		*Зачет*
		Оценка	*Буквенный эквивалент*	*Буквенный эквивалент*
86 - 100	Продвинутый	5	отлично	Зачтено
66 - 85	Повышенный	4	хорошо	
51 - 65	Пороговый	3	удовлетворительно	
0 - 50	Первый	2	неудовлетворительно	Не зачтено

5 Материально-техническое, программное и учебно-методическое обеспечение дисциплины.

5.1 Учебная литература

Основная учебная литература

Парфенов, Ю. П. Постреляционные хранилища данных : учебное пособие для вузов / Ю. П. Парфенов ; под научной редакцией Н. В. Папуловской. — Москва : Издательство Юрайт, 2023. — 121 с. — (Высшее образование). — ISBN 978-5-534-09837-2. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/514724.

Дополнительная учебная литература

Миркин, Б. Г. Введение в анализ данных : учебник и практикум / Б. Г. Миркин. — Москва : Издательство Юрайт, 2023. — 174 с. — (Высшее образование). — ISBN 978-5-9916-5009-0. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/511121.

5.2 Материально-техническое и программное обеспечение дисциплины.

Учебные занятия по дисциплине проводятся в учебных аудиториях КГПИ ФГБОУ ВО «КемГУ»:

610 Учебная аудитория (мультимедийная)для проведения: - занятий лекционного типа; - текущего контроля и промежуточной аттестации. **Специализированная (учебная) мебель:** доска меловая, кафедра, столы, стулья. **Оборудование для презентации учебного материала**: стационарное - компьютер, экран, проектор. **Используемое программное обеспечение:** MSWindows (MicrosoftImaginePremium 3 year по сублицензионному договору № 1212/КМР от 12.12.2018 г. до 12.12.2021 г.), LibreOffice (свободно распространяемое ПО), FoxitReader (свободно распространяемое ПО), Firefox 14 (свободно распространяемое ПО), Яндекс.Браузер (отечественное свободно распространяемое ПО). **Интернет с обеспечением доступа в ЭИОС**.	Учебный корпус №4. 654079, Кемеровская область, г. Новокузнецк, пр-кт Металлургов, д. 19
502 Компьютерный класс. Учебная аудитория (мультимедийная) для проведения: - занятий лекционного типа; - занятий семинарского (практического) типа; - занятий лабораторного типа; - групповых и индивидуальных консультаций; - самостоятельной работы; - текущего контроля и промежуточной аттестации. **Специализированная (учебная) мебель**: доска меловая, столы компьютерные, стулья. **Оборудование для презентации учебного материала**: *стационарное* - компьютер, экран, проектор, наушники. **Лабораторное оборудование**: стационарное – компьютеры для обучающихся (16 шт.). **Используемое программное обеспечение**: MSWindows (MicrosoftImaginePremium 3 year по сублицензионному договору № 1212/КМР от 12.12.2018 г. до 12.12.2021 г.), LibreOffice (свободно распространяемое ПО), Firefox 14 (свободно распространяемое ПО), Яндекс.Браузер (отечественное свободно распространяемое ПО), MicrosoftVisualStudio (MicrosoftImaginePremium 3 year по сублицензионному договору № 1212/КМРот 12.12.2018 г. до 12.12.2021 г.), Среда статистических вычислений Rv.4.0.2 (свободно распространяемое ПО). **Интернет с обеспечением доступа в ЭИОС.**	Учебный корпус №4. 654079, Кемеровская область, г. Новокузнецк, пр-кт Металлургов, д. 19

5.3 Современные профессиональные базы данных и информационные справочные системы.

Перечень СПБД и ИСС по дисциплине

CITForum.ru - on-line библиотека свободно доступных материалов по информационным технологиям на русском языке - http://citforum.ru

Научная электронная библиотека eLIBRARY.RU – крупнейший российский информационный портал в области науки, технологии, медицины и образования, содержащий рефераты и полные тексты - www.elibrary.ru

6 Иные сведения и (или) материалы.

6.1. Примерные вопросы и задания для промежуточной аттестации

Семестр 4

Таблица 9 - Примерные теоретические вопросы и практические задания к

экзамену

Теоретические вопросы

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 можно исследовать.
6. Понятия Web Mining и Web Analytics. Этапы аналитики в соответствии со стандартом 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. 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. Как обеспечивается отказоустойчивость Map-Reduce.
24. Технология Hadoop. MapReduce в Hadoop. Структура программы в Hadoop.
25. Хранилища Больших данных. Примеры распределенных хранилищ.
26.

Практические задания

1. Используя синтаксический анализ HTML в качестве способа извлечения информации с web-страниц, разработать программу по сбору информации методами Web-scrappinga и продемонстрировать результат ее работы.
2. Используя сопоставление текстовых шаблонов в качестве способа извлечения информации с web-страниц, разработать программу по сбору информации методами Web-scrappinga и продемонстрировать результат ее работы.
3. Используя метод нечеткого поиска в качестве способа извлечения информации с web-страниц, разработать программу по сбору информации методами Web-scrappinga и продемонстрировать результат ее работы.
4. Используя метод булева поиска в качестве способа извлечения информации с web-страниц, разработать программу по сбору информации методами Web-scrappinga и продемонстрировать результат ее работы.

Составитель (и): старший преподаватель кафедры МФММ Гаврилова Ю.С.
(*фамилия, инициалы и должность преподавателя (ей)*)