МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«ДАГЕСТАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
Факультет математики и компьютерных наук
Кафедра прикладной математики

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ
по дисциплине
«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

Фонд оценочных средств по дисциплине «Извлечение и анализ интернет данных» составлена в 2023 году в соответствии с требованиями ФГОС ВО бакалавриата по направлению подготовки 01.03.05 - статистика от 14.08.2020 г. № 1032

Разработчики:

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

Фонд оценочных средств по дисциплине «Извлечение и анализ интернет данных» одобрен:
на заседании кафедры Прикладной математики от «20» января 2023 г., протокол № 5

Зав. кафедрой_____Кадиев Р.И.

на заседании Методической комиссии факультета МиКН от «25»января 2023г., протокол №4.

Председатель_____Ризаев М.К.

Фонд оценочных средств «Извлечение и анализ интернет данных» согласован с учебно-методическим управлением

«20» февраля 2023 г _____

1. ПАСПОРТ
ФОНДА ОЦЕНОЧНЫХ СРЕДСТВ
по дисциплине
«Извлечение и анализ интернет данных»

1.1. Основные сведения о дисциплине

Общая трудоемкость дисциплины составляет 2 зачетные единицы (72 академических часов).

Вид работы	Трудоемкость, академических часов		
	6 семестр		всего
Общая трудоёмкость	108		108
Контактная работа:	48		48
Лекции (Л)	16		16
Практические занятия (ПЗ)	16		16
Лабораторные занятия (ЛЗ)	16		16
Консультации			
Промежуточная аттестация (зачет, экзамен)	зачет		
Самостоятельная работа			
1. работа с лекционным материалом, с учебной литературой	8		8
2. опережающая самостоятельная работа (изучение нового материала до его изложения на занятиях)	8		8
3. выполнение домашних заданий	8		8
4. подготовка к лабораторным работам, к практическим занятиям	8		8
5. подготовка к коллоквиуму	8		8
6. подготовка к контрольным работам	8		8
7. подготовка к зачету	12		12

1.2. Требования к результатам обучения по дисциплине, формы ихконтроля и виды оценочных средств

ПАСПОРТ ФОНДА ОЦЕНОЧНЫХ СРЕДСТВ
*по дисциплине «*Извлечение и анализ интернет данных*»*

№ п/п	Контролируемые модули, разделы (темы) дисциплины	Код контролируемой компетенции (или её части)	Оценочные средства		Способ контроля
			наименование	№№ заданий	
1	Модуль 1 **Аналитика в сети Интернет**	ПК-1	Вопросы для собеседования	1-18	устно
		ПК-1	Контрольные работы	1	письменно

2	Модуль 2 **Методология сбора данных из сетевых источников**	ПК-1	Лабораторные работы	1-3	письменно
		ПК-1	Вопросы для собеседования	10-33	устно
		ПК-1	Контрольные работы	2	письменно
		ПК-1	Лабораторные работы	4-6	письменно
3	Модуль 3 **Типы информационных систем. Устройство ипринцип работы поисковых систем.**	ПК-1	Вопросы для собеседования	34-46	устно
		ПК-1	Контрольные работы	3	письменно
		ПК-1	Лабораторные работы	7,8	письменно

1.3. Показатели и критерии определения уровня сформированности компетенций

№ п/п	Код компетенции	Уровни сформированности компетенции			
		Недостаточный	Удовлетворительный (достаточный)	Базовый	Повышенный
		Отсутствие признаков удовлетворительного уровня	Знать: Уметь: Владеть:	Знать: Уметь: Владеть:	Знать: Уметь: Владеть:
2	**ПК-1**	Не знает на достаточном уровне: стандартные методы и технические средства для статистических наблюдений. Не умеет на достаточном уровне применить стандартные методы и технические средства при статистических наблюдениях. Не владеет на достаточном уровне методами и техническими средствами для	Знает на достаточном уровне стандартные методы и технические средства для статистических наблюдений. Умеет на достаточном уровне применить стандартные методы и технические средства при статистических наблюдениях. Владеет на достаточном уровне методами и техническими	Знает на хорошем уровне стандартные методы и технические средства для статистических наблюдений. Умеет на хорошем уровне применить стандартные методы и технические средства при статистических наблюдениях. Владеет на хорошем уровне методами и техническими	Знает в совершенстве стандартные методы и технические средства для статистических наблюдений. Умеет в совершенстве применить стандартные методы и технические средства при статистических наблюдениях. Владеет в совершенстве методами и техническими средствами для

		статистических наблюдений.	средствами для статистических наблюдений	средствами для статистических наблюдений	статистически х наблюдений

2. КОНТРОЛЬНЫЕ ЗАДАНИЯ И ИНЫЕ МАТЕРИАЛЫ ОЦЕНКИ знаний, умений, навыков и (или) опыта деятельности, характеризующие этапы формирования компетенций в процессе освоения дисциплины «Извлечение и анализ интернет данных»

Контрольные работы

Контрольная работа 1

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.

Контрольная работа 2

1. Deep WEB. Какие ресурсы его составляют. Какими средствами его можно исследовать.
2. Понятия Web Mining и Web Analytics. Этапы аналитики в соответствии со стандартом CRISP-DM.
3. Задачи Data Mining. Направления Data Mining.
4. Понятие и задачи Web Content Mining.

Контрольная работа 3

1. Как работают алгоритмы индексирования. Необходимость ранжирования и задачи машинного обучения в приложении к информационному поиску.
2. Охарактеризуйте модели информационного поиска.
3. Изложите подробно принцип булевой модели информационного поиска (ИП), возможные средства оптимизации запроса.

Критерии оценки:

- оценка «отлично» выставляется студенту, если верно и правильно выполнено 90%-100% заданий;
- оценка «хорошо» выставляется студенту, если верно и правильно выполнено 70%-80% заданий;
- оценка «удовлетворительно» выставляется студенту, если верно и правильно решено 50%-60% заданий, возможны некоторые исправления при решении;
- оценка «неудовлетворительно» выставляется студенту, если верно выполнено менее 50% заданий;

Вопросы для коллоквиумов, собеседования

Модуль 1. Аналитика в 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.
8. Невидимый WEB, его возможности и характеристики.
9. Инструменты и технологии работы в невидимом WEB.
10. Проблемы, возникающие при поддержании актуальности информации на сайте.
11. Определение CMS. Краткое описание CMS.
12. Динамический и статический сайты.
13. Характеристика контента. Создание контента.
14. Управление автоматизированными деловыми процессами.
15. Распространение контента.
16. Персонализация и глобализация контента. Критерии классификации систем управления контентом.
17. Простая CMS. Шаблонная CMS. Профессиональная CMS. Универсальная CMS.
18. 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

19. Определение понятий WEB Mining и Data Mining.
20. Отличия между ними. Задачи и этапы извлечения знаний из WEB.
21. Направления WEB-mining: Извлечение Web-контента (Web Content Mining);
22. Извлечение Web-структур (Web Structure Mining);
23. Исследование использования Web-ресурсов (Web Usage Mining)
24. Понятие бизнес- аналитического решения.
25. Анализ журнала посещаемости сайта.
26. Заказные статистические исследования.
27. Определение профиля сайта.
28. Определение перечня сайтов, посещаемых вашей аудиторией.
29. Определение целевой аудитории сайта. Типы посетителей сайтов.
30. Модели поведения посетителей сайта. Пользователи Интернет магазинов.
31. Булева модель, векторная модель, вероятностная модель, гибридная модель.
32. Математические особенности обработки информации разными моделями.
33. Сферы их применения.

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

34. Системы переработки информации.
35. Типы информационных систем. Уточнение структуры информационных систем.
36. Информационные системы Интернета.
37. Понятие поисковой системы. Принципы работы поисковых систем, которые нужно учитывать при продвижении сайта.
38. Виды поисковых роботов.
39. Порядок индексации сайтов. Порядок поисковой выдачи.
40. Принципы алгоритмов выдачи поисковой системы Яндекс и Google.
41. Выбор ключевых слов для продвижения сайта. Типы запросов по частотности. Типы запросов по степени конверсии.
42. Понятие семантического ядра. Создание семантического ядра. Выбор ключевых

страниц сайта. Распределение семантического ядра.
43. Анализ сайтов конкурентов.
44. Расчет сложности продвижения сайта. Выбор основной стратегии поискового продвижения сайта.
45. Требования к хранилищам данных, OLTP и OLAP системы.
46. Нереляционные базы данных.

Критерии оценки:

- оценка «отлично» выставляется студенту, если изложение полученных знаний в устной форме полное, в системе, в соответствии с требованиями учебной программы; допускаются единичные несущественные ошибки, самостоятельно исправляемые учащимися;
- оценка «хорошо» выставляется студенту, если изложение полученных знаний в устной форме полное, в системе, в соответствии с требованиями учебной программы; допускаются, отдельные несущественные ошибки, исправляемые учащимися после указания преподавателя на них;
- оценка «удовлетворительно» выставляется студенту, если изложение полученных знаний неполное, однако это не препятствует усвоению последующего программного материала; допускаются отдельные существенные ошибки, исправляемые с помощью преподавателя;
- оценка «неудовлетворительно» выставляется студенту, если изложение учебного материала неполное, бессистемное, что препятствует усвоению последующей учебной информации; существенные ошибки, не исправляемые даже с помощью преподавателя;

Комплект тестовых заданий для контроля

1. 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

Критерии оценки:

- оценка «отлично» выставляется студенту, если верно и правильно выполнено 90%-100% заданий;
- оценка «хорошо» выставляется студенту, если верно и правильно выполнено 70%-80% заданий;
- оценка «удовлетворительно» выставляется студенту, если верно и правильно решено 50%-60% заданий, возможны некоторые исправления при решении;
- оценка «неудовлетворительно» выставляется студенту, если верно выполнено менее 50% заданий;

Лабораторные работы

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, выполнение поиска 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 данных с 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*

Критерии оценки:

- оценка «отлично» выставляется студенту, если выполнены все задания лабораторной работы, составлен отчет по работе
- оценка «хорошо» выставляется студенту, если выполнены почти все задания, за исключением отдельных пунктов, лабораторной работы, составлен отчет по работе
- оценка «удовлетворительно» выставляется студенту, если выполнены больше половины заданий лабораторной работы, составлен отчет по работе
- оценка «неудовлетворительно» выставляется студенту, если выполнены меньше половины заданий лабораторной работы и не составлен отчет по работе

Темы эссе (рефератов, докладов, сообщений)

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 анализ данных Московской биржи

Реферат оценивается следующим образом:

- соответствие содержания теме- 4 балла;
- глубина проработки материала, 3 балла;
- грамотность и полнота использования источников, 1 балл;
- соответствие оформления реферата требованиям, 2 балла;
- доклад, 5 баллов;
- умение вести дискуссию и ответы на вопросы, 5 баллов.

Максимальное количество баллов: 20.

Критерии оценки:

- оценка «отлично» выставляется студенту, если набрал 19-20 баллов;
- оценка «хорошо» выставляется студенту, если набрал 15-18 баллов;
- оценка «удовлетворительно» выставляется студенту, если набрал 10-14 баллов;
- оценка «неудовлетворительно» выставляется студенту, если набрал менее 10 баллов;

Вопросы 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 кластерногоанализа 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.

Критерии оценки:

- «зачтено» выставляется студенту, если изложение полученных знаний в устной форме полное, в системе, в соответствии с требованиями учебной программы; допускаются, отдельные несущественные ошибки, исправляемые учащимися после указания преподавателя на них;
- «не зачтено» выставляется студенту, если изложение учебного материала неполное, бессистемное, что препятствует усвоению последующей учебной информации; существенные ошибки, не исправляемые даже с помощью преподавателя.

Рекомендуемые границы оценок:
«зачтено» - не менее 51% правильных ответов,
«не зачтено»- менее 51% правильных ответов.