

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение  
высшего образования  
«Луганский государственный университет имени Владимира Даля»

Экономический факультет  
Кафедра экономической кибернетики и прикладной статистики

УТВЕРЖДАЮ:  
Декан экономического факультета  
Тхор Е.С.  
(подпись)  
« 24 » \_\_\_\_\_ 2023 года



**РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ**

**«МЕТОДЫ ПОИСКА СТРУКТУРИРОВАННОЙ  
И НЕСТРУКТУРИРОВАННОЙ ИНФОРМАЦИИ»**

По направлению подготовки 38.03.05 Бизнес-информатика  
Профиль: «Экономическая аналитика и бизнес-статистика»

Луганск – 2023

Лист согласования РПУД


Рабочая программа учебной дисциплины «Методы поиска структурированной и неструктурированной информации» по направлению подготовки 38.03.05 Бизнес-информатика. – 38 с.

Рабочая программа учебной дисциплины «Методы поиска структурированной и неструктурированной информации» составлена с учетом Федерального государственного образовательного стандарта высшего образования по направлению подготовки 38.03.05 Бизнес-информатика, утвержденного приказом Министерства науки и высшего образования Российской Федерации от 29 июня 2020 года № 838.

СОСТАВИТЕЛЬ (СОСТАВИТЕЛИ):

к.э.н., доцент Спорняк С.А.

Рабочая программа дисциплины утверждена на заседании кафедры экономической кибернетики и прикладной статистики «18» 04 2023 г., протокол № 26

Заведующий кафедрой экономической кибернетики  
и прикладной статистики  А.В. Велигура

Переутверждена: «  » \_\_\_\_\_ 20   г., протокол № \_\_\_\_\_

Согласована (для обеспечивающей кафедры):

Декан экономического факультета  Тхор Е.С.

Переутверждена: «  » \_\_\_\_\_ 20   года, протокол № \_\_\_\_\_

Рекомендована на заседании учебно-методической комиссии экономического факультета «21» апреля 2023 г., протокол № 4.

Председатель учебно-методической  
комиссии экономического факультета  Е.Н. Шаповалова

## Структура и содержание дисциплины

### 1. Цели и задачи дисциплины, ее место в учебном процессе

Цель изучения дисциплины – формирование знания, практических навыков и умений поиска экономической информации в глобальной сети Интернет, информационных банках и массивах; обработка информации с помощью офисных инструментальных средств и технологий.

Задачи:

- формирование знания, практических навыков и умений поиска экономической информации в глобальной сети Интернет, информационных банках и массивах;
- обработка информации с помощью офисных инструментальных средств и технологий.

### 2. Место дисциплины в структуре ОПОП ВО. Требования к результатам освоения содержания дисциплины

Дисциплина «Методы поиска структурированной и неструктурированной информации» относится к части, формируемой участниками образовательных отношений.

Необходимыми условиями для освоения дисциплины являются:

**знания:**

- сущности и значения информации в развитии современного общества, основные закономерности создания и функционирования информационных процессов в финансово-экономической сфере;
- методов и технологий поиска и обработки экономической информации средствами Интернета и офисных приложений;
- основных источников экономической информации.

**умения:**

- работы с информацией в глобальных компьютерных сетях;
- применять при решении прикладных финансово-экономических задач современные информационные технологии для поиска и обработки информации в системе глобальных информационных ресурсов;
- готовить аналитические обзоры, отчеты и презентации на основе найденной информации;
- использовать полученные знания, навыки и умения для формирования и развития профессиональных компетенций;

**навыки:**

- основных методов, способов и средств поиска, получения, хранения и переработки экономической информации.

Предшествующими дисциплинами, формирующими начальные знания, являются следующие: «Бизнес-информатика», «Бизнес-информатика 2», «Экономическая статистика», «Математические методы принятия решений», «Математика», «Исследование операций», «Статистика», «Эконометрика», «Анализ данных средствами языка программирования R», «Многомерные статистические методы», и служит основой для освоения дисциплин

«Прогнозирование социально-экономических процессов», «Статистика производства и рынка товаров и услуг» и преддипломной практики.

### 3. Требования к результатам освоения содержания дисциплины

Код и наименование компетенции	Индикаторы достижений компетенции (по реализуемой дисциплине)	Перечень планируемых результатов
<p>ПК-5. Способен проводить статистическое наблюдение и группировку, в том числе на основании данных статистических регистров с использованием стандартных методик и технических средств</p>	<p>ПК-5.1. Использует источники, основные способы сбора, поиска и систематизации статистической информации, приемы структурирования исходных данных с применением информационно-коммуникационных технологий</p>	<p><b>Знать:</b>                      сущность и значение информации в развитии современного общества, методы и технологии поиска и обработки экономической информации средствами Интернета и офисных приложений;                      основные источники информации. ключевые принципы работы с ПК, методов сбора и обработки первичной и вторичной информации из различных источников, в том числе сети Интернет;                      процесс сбора финансовой, экономической, статистической и бухгалтерской информации;                      возможность обработки собранной информации при помощи информационных технологий;                      основные источники информации при подготовке аналитического отчета и информационного обзора;                      основных методов решения аналитических и исследовательских задач.</p>
		<p><b>Уметь:</b>                      работать с информацией в глобальных компьютерных сетях;                      применять при решении прикладных финансово-экономических задач современные информационные технологии для поиска и обработки информации в системе глобальных информационных ресурсов;                      готовить аналитические обзоры, отчеты и презентации на основе найденной информации;</p>
		<p><b>Владеть:</b>                      основными методами, способами средств поиска, получения, хранения и переработки экономической информации.</p>

## 4. Структура и содержание дисциплины

### 4.1. Объем учебной дисциплины и виды учебной работы

Вид учебной работы	Объем часов (зач. ед.)		
	Очная форма	Очно-заочная форма	Заочная форма
Общая учебная нагрузка (всего)	108 (3 зач. ед)	108 (3 зач. ед)	108 (3 зач. ед)
Обязательная контактная работа (всего) в том числе:	56	28	12
Лекции	28	14	6
Семинарские занятия	-	-	-
Практические занятия	28	14	6
Лабораторные работы	-	-	-
Курсовая работа (курсовой проект)	-	-	-
Другие формы и методы организации образовательного процесса ( <i>расчетно-графические работы, индивидуальные задания и т.п.</i> )	-	-	-
Самостоятельная работа студента (всего)	52	80	96
Форма аттестации	зачет с оценкой	зачет с оценкой	зачет с оценкой

### 4.2. Содержание разделов дисциплины

#### ***Тема 1. ПОНЯТИЕ ИНФОРМАЦИИ И ИНФОРМАЦИОННОГО ПОИСКА***

Понятие информации. Характеристики дискретных источников информации. Свойства информации. Условная информация. Формы адекватности информации. Качества информации. Структурированная и неструктурированная информация.

Понятие информационного поиска. Основные категории информационного поиска: документ, слово, термин, запрос, релевантность, полнота и точность. Этапы информационного поиска. Виды информационного поиска. Методы поиска. Эффективность информационного поиска.

Современные поисковые средства поиска информации. Поисковые системы, каталоги, мета-поисковые системы. Инструменты поиска: язык запросов поисковиков, лингвистические особенности языка разыскиваемых инструментов.

#### ***Тема 2. БУЛЕВ ПОИСК***

Прямой поиск и поиск по индексу. Инвертированный индекс. Обработка булевых запросов. Сравнение расширенной булевой модели и ранжированного поиска.

#### ***Тема 3. ЛЕКСИКОН И СПИСКИ СЛОВОПОЗИЦИЙ***

Схематизация документа и декодирование последовательности символов. Определение лексикона терминов. Быстрое пересечение

инвертированных списков с помощью указателей пропусков. Словопозиции с координатами и фразовые запросы. Лемматизация, морфологический анализ. Принципы работы морфологического анализатора. Процедурный, табличный и вероятностный подходы. Примеры библиотек.

#### ***Тема 4. СЛОВАРИ И НЕЧЕТКИЙ ПОИСК***

Поисковые структуры для словарей. Запросы с джокером. Исправление опечаток. Фонетические исправления.

#### ***Тема 5. ПОСТРОЕНИЕ ИНДЕКСА***

Характеристики аппаратного обеспечения. Блочное индексирование, основанное на сортировке. Однопроходное индексирование в оперативной памяти. Распределенное индексирование. Динамическое индексирование. Другие типы индексов.

#### ***Тема 6. СЖАТИЕ ИНДЕКСОВ***

Статистические характеристики терминов, документов и коллекций. Закон Ципфа. Сжатие словаря. Сжатие инвертированного файла. Байтовое кодирование переменной длины. Гамма-коды.

#### ***Тема 7. РАНЖИРОВАНИЕ, ВЗВЕШИВАНИЕ ТЕРМИНОВ И МОДЕЛЬ ВЕКТОРНОГО ПРОСТРАНСТВА***

Параметрические и зонные индексы. Частота термина и взвешивание. Модель векторного пространства для ранжирования. Варианты функций  $TF*IDF$ .

#### ***Тема 8. РАНЖИРОВАНИЕ В ПОЛНОФУНКЦИОНАЛЬНОЙ ПОИСКОВОЙ СИСТЕМЕ***

Эффективное ранжирование. Компоненты информационно-поисковой системы. Влияние операторов языка запросов на ранжирование в векторном пространстве.

#### ***Тема 9. ОЦЕНКА СИСТЕМ ИНФОРМАЦИОННОГО ПОИСКА***

Роль и сложность оценки в информационном поиске. Стандартные тестовые коллекции. Оценка неранжированных результатов поиска. Оценка ранжированных результатов поиска. Оценка релевантности. Качество системы и ее полезность для пользователя. С니ппеты.

#### ***Тема 10. ОБРАТНАЯ СВЯЗЬ ПО РЕЛЕВАНТНОСТИ И РАСШИРЕНИЕ ЗАПРОСА***

Обратная связь по релевантности и псевдорелевантности. Глобальные методы переформулировки запросов.

### ***Тема 11. ОСНОВЫ ВЕБ-ПОИСКА***

История. Характеристики веба. Реклама как экономическая модель. Оценка размера индекса поисковых машин веба. Нечеткие дубликаты документов и модель шинглов. Обход веба. Распределенные индексы. Анализ ссылок. Алгоритм.

### ***Тема 12. КОНЦЕПЦИЯ «БОЛЬШИХ ДАННЫХ»***

Что такое «Большие данные», и что они нам сулят. Разница между бизнес-аналитикой и «Большими данными». Устаревание информации. Рост объемов данных на фоне вытеснения аналоговых средств хранения. Корректная интерпретация информационных потоков. Обработка информационных потоков. Предпосылки применения контент-анализа в различных исследованиях. Необходимость в аналитической работе с большими данными. Явная (выраженная) и скрытая (структурная) информация. Количественная и качественная стратегия анализа текстов. Возможности и ограничения каждого из подходов. Процедура контент-анализа. Определение круга проблем для контент-анализа. Начальный этап исследования: формулирование целей и задач исследования, выбор эмпирического материала, выдвижение рабочих гипотез. Операциональный этап исследования: определение категорий и подкатегорий, выбор единиц анализа, установление правил кодирования. Этап счета. Этап интерпретации результатов. Презентация результатов. Типичные ошибки при проведении контент-анализа. Технические признаки, характеризующие «Большие данные». Принцип V3 – Volume (объем данных), Variety (разнообразие данных) и Velocity (скорость генерации и работы с данными). Интеграция, миграция и построение хранилищ данных. Высокопроизводительные вычисления (High Performance Computing, HPC) при выполнении аналитических исследований. Grid computing (распределенные вычисления на нескольких серверах), in-database analytics (частичный перевод нагрузки при аналитических вычислениях в СУБД, а также регламентное применение готовых аналитических моделей к новым данным полностью на стороне СУБД) и in-memory analytics (применение аналитики прямо в оперативной памяти сервера СУБД).

### ***Тема 13. НЕСТРУКТУРИРОВАННАЯ ИНФОРМАЦИЯ***

Эвристические алгоритмы поиска, эволюционное вычисление, этапы генетического алгоритма: задание целевой функции (приспособленности) для особей популяции, создание начальной популяции, размножение (скрещивание), мутирование, вычисление значения целевой функции для всех особей, формирование нового поколения (селекция). Задача кластеризации, методы кластеризации, иерархическая кластеризация, алгоритм k-средних, зонтичная кластеризация, методы ненаправленного обучения (Unsupervised Learning). Постановка задачи классификации, подходы и применения, построение и обучение классификатора, оценка качества классификации, рубрикации тренировочных данных (Training Data

Set), методы управляемого (направляемого) обучения (Supervised Learning). Методы распознавания образов, дискриминантный анализ, нелинейная оптимизация, этапы формирования нейронных сетей: сбор данных для обучения, подготовка и нормализация данных, выбор топологии сети, экспериментальный подбор характеристик сети, экспериментальный подбор параметров обучения, собственно обучение, проверка адекватности обучения, корректировка параметров, окончательное обучение, вербализация сети с целью дальнейшего использования. Совместное использование компьютерных технологий и лингвистики для создания алгоритмов, позволяющих анализировать естественные (человеческие) языки. Применение методов обработки естественных языков и других аналитических методов для выявления и извлечения из анализируемого текста субъективной информации, характеризующей настроения, мнения, отношение людей к проблеме. Рассмотрение следующих основных задач: синтез речи, распознавание речи, анализ текста, синтез текста, машинный перевод, вопросно-ответные системы, информационный поиск, извлечение информации, анализ тональности текста, анализ высказываний, упрощение текста.

#### ***Тема 14. АППАРАТНОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ «БОЛЬШИХ ДАННЫХ»***

Вычисления некоторых наборов распределенных задач с использованием большого количества компьютеров, образующих кластер. Шаги Map и Reduce. Предварительная обработка входных данных и свёртка данных. Концепция параллелизма. Шаблоны доступа к данным, хеш-таблица, деревья, таксономия NoSQL, колоночные СУБД, bigtable. Разработка и выполнение распределённых программ, расширение вычислительных мощностей посредством добавления в кластер дополнительных узлов, технология Hadoop, распределённая файловая система HDFS (Hadoop Distributed File System), интеграция с NoSQL и MapReduce.

#### ***Тема 15. МАСШТАБИРОВАНИЕ И МНОГОУРОВНЕВОЕ ХРАНЕНИЕ «БОЛЬШИХ ДАННЫХ»***

Модели развёртывания: частное облако, публичное облако, гибридное облако, общественное облако. Модели обслуживания: программное обеспечение, платформа, инфраструктура. Экономические аспекты центров обработки данных. Безопасность при хранении и пересылке данных. Проблема «последней мили». Обработка Fast Data, подтверждение и корректировка априорных знаний и гипотез, синхронизация скорости работы с ростом объема данных. Получение знаний посредством Big Analytics, преобразования зафиксированной в данных информации в новое знание, принцип «обучения с учителем». Высший уровень работы с данными Deep Insight, обучение без учителя (unsupervised learning), использование современных методов аналитики, а также различные способы визуализации, обнаружение знаний и закономерностей, априорно неизвестных.



## ***Тема 16. ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ «БОЛЬШИХ ДАННЫХ»***

Практическое применение решений IBM Cognos Analytics и ресурсов платформы IBM Bluemix. Понятие шаблона, создание правил и категорий. Персональная база данных, фразовый поиск, нечеткий поиск. Возможности уточнения результатов запросов с учетом структуры текста. Анализ совместной встречаемости (collocate analysis) и коэффициент связи категорий (Z-score). Практическое применение решений IBM Cognos Analytics и ресурсов платформы IBM Bluemix. Контент-анализ массовой корреспонденции и социологических опросов. Прямые пропорциональные закономерности, аддитивные закономерности, мультипликативные закономерности.

## ***Тема 17. ВЕДЕНИЕ В АНАЛИЗ БОЛЬШИХ ДАННЫХ. ОБЗОР ИСТОЧНИКОВ ИНФОРМАЦИИ***

Основные определения, термины, задачи анализа больших данных. Вопросы безопасности. Понятие Data Mining. Когнитивный анализ данных. Обзор источников информации для Big Data (открытые источники информации: статистические сборники, опубликованные отчеты и результаты исследований; доступ к закрытой информации). Методики сбора данных.

## ***Тема 18. ТЕХНОЛОГИИ ХРАНЕНИЯ И ОБРАБОТКИ БОЛЬШИХ ДАННЫХ***

Обзор технологий хранения больших данных. Базы данных. Системы управления базами данных. Модели данных. Подготовка исходных данных для анализа: первичная обработка и визуализация имеющихся данных.

## ***Тема 19. СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ДАННЫХ***

Содержание темы. Основные понятия математической статистики. Методы анализа данных: дескриптивная статистика, параметрические, непараметрические, номинальные методы (корреляционный, регрессионный, дисперсионный анализы, кластерный, дискриминантный, факторный анализы).

## ***Тема 20. СОВРЕМЕННЫЕ ПРОГРАММНЫЕ СРЕДСТВА АНАЛИЗА БОЛЬШИХ ОБЪЕМОВ ИНФОРМАЦИИ***

Обзор современных популярных программных средства анализа данных: Statistica, SPSS, Excel, R-Studio и другие; их преимущества и недостатки.

### 4.3. Лекции

№ п/п	Название темы	Объем часов		
		Очная форма	Очно-заочная форма	Заочная форма
1	Понятие информации и информационного поиска	1	0,5	0,3
2	Булев поиск	1	0,5	0,3
3	Лексикон и списки словопозиций	1	0,5	0,3
4	Словари и нечеткий поиск	1	0,5	0,3
5	Построение индекса	1	0,5	0,3
6	Сжатие индексов	1	0,5	0,3
7	Ранжирование, взвешивание терминов и модель векторного пространства	1	0,5	0,3
8	Ранжирование в полнофункциональной начальной поисковой системе	1	0,5	0,3
9	Оценка систем информационного поиска	1	0,5	0,3
10	Обратная связь по релевантности и расширение запроса	1	0,5	0,3
11	Основы веб-поиска	1	0,5	0,3
12	Концепция «Больших Данных»	1	0,5	0,3
13	Неструктурированная информация	2	1	0,3
14	Аппаратное и программное обеспечение «Больших Данных»	2	1	0,3
15	Масштабирование и многоуровневое хранение «Больших Данных»	2	1	0,3
16	Практическое применение «Больших Данных»	2	1	0,3
17	Ведение в анализ больших данных. Обзор источников информации	2	1	0,3
18	Технологии хранения и обработки больших данных	2	1	0,3
19	Статистические методы анализа данных	2	1	0,3
20	Современные программные средства анализа больших объемов информации	2	1	0,3
<b>Итого:</b>		<b>28</b>	<b>14</b>	<b>6</b>

### 4.4. Практические (семинарские) занятия

№ п/п	Название темы	Объем часов		
		Очная форма	Очно-заочная форма	Заочная форма
1	Понятие информации и информационного поиска	1	0,5	0,3
2	Булев поиск	1	0,5	0,3
3	Лексикон и списки словопозиций	1	0,5	0,3
4	Словари и нечеткий поиск	1	0,5	0,3
5	Построение индекса	1	0,5	0,3
6	Сжатие индексов	1	0,5	0,3
7	Ранжирование, взвешивание терминов и модель векторного пространства	1	0,5	0,3
8	Ранжирование в полнофункциональной начальной поисковой системе	1	0,5	0,3
9	Оценка систем информационного поиска	1	0,5	0,3

10	Обратная связь по релевантности и расширение запроса	1	0,5	0,3
11	Основы веб-поиска	1	0,5	0,3
12	Концепция «Больших Данных»	1	0,5	0,3
13	Неструктурированная информация	2	1	0,3
14	Аппаратное и программное обеспечение «Больших Данных»	2	1	0,3
15	Масштабирование и многоуровневое хранение «Больших Данных»	2	1	0,3
16	Практическое применение «Больших Данных»	2	1	0,3
17	Ведение в анализ больших данных. Обзор источников информации	2	1	0,3
18	Технологии хранения и обработки больших данных	2	1	0,3
19	Статистические методы анализа данных	2	1	0,3
20	Современные программные средства анализа больших объемов информации	2	1	0,3
<b>Итого:</b>		<b>28</b>	<b>14</b>	<b>6</b>

**4.5. Лабораторные работы по дисциплине «Методы поиска структурированной и неструктурированной информации» не предполагаются учебным планом.**

#### **4.6. Самостоятельная работа студентов**

№ п/п	Название темы	Вид СРС	Объем часов		
			Очная форма	Очно-заочная форма	Заочная форма
1	Понятие информации и информационного поиска	Подготовка к практическому занятию	2	3	4
2	Булев поиск	Подготовка к практическому занятию	2	3	4
3	Лексикон и списки словопозиций	Подготовка к практическому занятию	2	3	4
4	Словари и нечеткий поиск	Подготовка к практическому занятию	2	3	4
5	Построение индекса	Подготовка к практическому занятию	2	4	4
6	Сжатие индексов	Подготовка к практическому занятию	2	4	4
7	Ранжирование, взвешивание терминов и модель векторного пространства	Подготовка к практическому занятию	2	4	4
8	Ранжирование в полнофункциональной начальной поисковой системе	Подготовка к практическому занятию	2	4	4

9	Оценка систем информационного поиска	Подготовка к практическому занятию	2	4	5
10	Обратная связь по релевантности и расширение запроса	Подготовка к практическому занятию	2	4	5
11	Основы веб-поиска	Подготовка к практическому занятию	2	4	5
12	Концепция «Больших Данных»	Подготовка к практическому занятию	2	4	5
13	Неструктурированная информация	Подготовка к практическому занятию	3	4	5
14	Аппаратное и программное обеспечение «Больших Данных»	Подготовка к практическому занятию	3	4	5
15	Масштабирование и многоуровневое хранение «Больших Данных»	Подготовка к практическому занятию	3	4	5
16	Практическое применение «Больших Данных»	Подготовка к практическому занятию	3	4	5
17	Ведение в анализ больших данных. Обзор источников информации	Подготовка к практическому занятию	3	4	5
18	Технологии хранения и обработки больших данных	Подготовка к практическому занятию	3	4	5
19	Статистические методы анализа данных	Подготовка к практическому занятию	3	4	5
20	Современные программные средства анализа больших объемов информации	Подготовка к практическому занятию	3	4	5
21	Зачет с оценкой	Подготовка к зачету с оценкой	4	4	4
<b>Итого:</b>			<b>52</b>	<b>80</b>	<b>96</b>

**4.7. Курсовые работы/проекты по дисциплине «Методы поиска структурированной и неструктурированной информации» не предполагаются учебным планом.**

## **5. Образовательные технологии**

Преподавание дисциплины ведется с применением следующих видов образовательных технологий: объяснительно-иллюстративного обучения (технология поддерживающего обучения, технология проведения учебной

дискуссии), информационных технологий (презентационные материалы), развивающих и инновационных образовательных технологий.

Практические занятия проводятся с использованием развивающих, проблемных, проектных, информационных (использование электронных образовательных ресурсов (электронный конспект) образовательных технологий.

## 6. Формы контроля освоения дисциплины

Текущая аттестация студентов производится в дискретные временные интервалы лектором и преподавателем(ями), ведущими практические занятия по дисциплине в следующих формах:

- вопросы для обсуждения на практических занятиях (устный опрос);
- контрольные работы (по вариантам);
- тесты.

Промежуточная аттестации по результатам освоения дисциплины проходит в форме устного зачета (включает в себя ответы на теоретические вопросы и ответы на тестовые задания). Студенты, выполнившие 75% текущих и контрольных мероприятий на «отлично», а остальные 25 % на «хорошо», имеют право на получение итоговой оценки.

В экзаменационную ведомость и зачетную книжку выставляются оценки по шкале, приведенной в таблице.

Шкала оценивания (экзамен)	Характеристика знания предмета и ответов	Зачеты
отлично (5)	Студент глубоко и в полном объеме владеет программным материалом. Грамотно, исчерпывающе и логично его излагает в устной или письменной форме. При этом знает рекомендованную литературу, проявляет творческий подход в ответах на вопросы и правильно обосновывает принятые решения, хорошо владеет умениями и навыками при выполнении практических задач.	зачтено
хорошо (4)	Студент знает программный материал, грамотно и по сути излагает его в устной или письменной форме, допуская незначительные неточности в утверждениях, трактовках, определениях и категориях или незначительное количество ошибок. При этом владеет необходимыми умениями и навыками при выполнении практических задач.	
удовлетворительно (3)	Студент знает только основной программный материал, допускает неточности, недостаточно четкие формулировки, непоследовательность в ответах, излагаемых в устной или письменной форме. При этом недостаточно владеет умениями и навыками при выполнении практических задач. Допускает до 30% ошибок в излагаемых ответах.	
неудовлетворительно (2)	Студент не знает значительной части программного материала. При этом допускает принципиальные ошибки в доказательствах, в трактовке понятий и категорий, проявляет низкую культуру знаний, не владеет основными умениями и навыками при выполнении практических задач. Студент отказывается от ответов на дополнительные вопросы.	не зачтено

## **7. Учебно-методическое и информационное обеспечение дисциплины:**

### **а) основная литература:**

1. Бессмертный, И.А. Системы искусственного интеллекта: учебное пособие для академического бакалавриата / И. А. Бессмертный. – 2-е изд., испр. и доп. – Москва: Издательство Юрайт, 2019. — 157 с. Режим доступа: <https://biblioonline.ru/bcode/423120>

2. Маннинг, Кристофер Д. Введение в информационный поиск: [пер. с англ.]/ Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце. — М. [и др.]: Вильямс, 2011. — 520 с.: ил. — ISBN 978-5-8459-1623-5. Английский вариант доступен: <https://nlp.stanford.edu/IR-book/>

3. Кожаринов А.С., Моделирование и анализ информационных и бизнес-процессов в информационных системах: метод. указ. к выполнению курсовых работ / А.С. Кожаринов. - М.: МИСиС, 2017. - 27 с. - Текст: электронный // ЭБС "Консультант студента": [сайт]. - URL: [http://www.studentlibrary.ru/book/Misis\\_362.html](http://www.studentlibrary.ru/book/Misis_362.html)

4. Чубукова И.А., Data Mining / Чубукова И.А. - М.: Национальный Открытый Университет "ИНТУИТ", 2016. (Основы информационных технологий) - ISBN 978-5-94774-819-2 - Текст: электронный // ЭБС "Консультант студента": [сайт]. - URL: <http://www.studentlibrary.ru/book/ISBN9785947748192.html>

5. Мельниченко А.С., Математическая статистика и анализ данных: учеб. пособие / А.С. Мельниченко - М.: МИСиС, 2018. - 45 с. - ISBN 978-5-906953-62-9 - Текст: электронный // ЭБС "Консультант студента": [сайт]. - URL: <http://www.studentlibrary.ru/book/ISBN9785906953629.html>

### **б) дополнительная литература:**

1. Афонин П.Н., Статистический анализ с применением современных программных средств: учебное пособие / Афонин П.Н., Афонин Д.Н. - СПб.: ИЦ Интермедия, 2017. - 100 с. - ISBN 978-4383-0080-9 - Текст: электронный // ЭБС "Консультант студента": [сайт]. - URL: <http://www.studentlibrary.ru/book/ISBN978438300809.html>

2. Кук Д., Машинное обучение с использованием библиотеки H2O / Кук Д. - М.: ДМК Пресс, 2018. - 250 с. - ISBN 978-5-97060-508-0 - Текст: электронный // ЭБС "Консультант студента": [сайт]. - URL: <http://www.studentlibrary.ru/book/ISBN9785970605080.html>

3. Джеймс Г., Введение в статистическое обучение с примерами на языке R / Джеймс Г., Уиттон Д., Хасты Т., Тибширани Р. - М.: ДМК Пресс, 2017. - 456 с. - ISBN 978-5-97060-495-3 - Текст: электронный // ЭБС "Консультант студента": [сайт]. - URL: <http://www.studentlibrary.ru/book/ISBN9785970604953.html>

4. Тагиева Р.Ф., Обработка экспериментальных данных. Ч.1: учебное пособие: в 2 ч. / Р.Ф. Тагиева, А.Н. Титов - Казань: Издательство КНИТУ, 2017. - 96 с. - ISBN 978-5-7882-2261-5 - Текст: электронный // ЭБС "Консультант студента": [сайт]. - URL: <http://www.studentlibrary.ru/book/ISBN9785788222615.html>

#### **в) методические рекомендации:**

1. Методические указания к практическим занятиям по дисциплине «Методы поиска структурированной и неструктурированной информации» для студентов направления подготовки 38.03.05 – Бизнес-информатика [Электронный ресурс] / сост. Ю.Е. Щеглов. – Луганск: ЛНУ им. В. Даля, 2019. – 78 с.

2. Методические указания к самостоятельной работе по дисциплине «Методы поиска структурированной и неструктурированной информации» для студентов направления подготовки 38.03.05 – Бизнес-информатика [Электронный ресурс] / сост. Ю.Е. Щеглов. – Луганск: ЛНУ им. В. Даля, 2019. – 26 с.

#### **г) Интернет-ресурсы:**

1. Министерство образования и науки Российской Федерации. – <http://минобрнауки.рф/>

2. Федеральная служба по надзору в сфере образования и науки. – <http://obrnadzor.gov.ru/>

3. Федеральная служба государственной статистики. – <https://rosstat.gov.ru/>

4. Территориальный орган Федеральной службы государственной статистики по Луганской Народной Республике – <https://www.gkslnr.su/>

5. Портал Федеральных государственных образовательных стандартов высшего образования – <http://fgosvo.ru/>

6. Федеральный портал «Российское образование» – <http://www.edu.ru/>

7. Информационная система «Единое окно доступа к образовательным ресурсам» – <http://window.edu.ru/>

8. Федеральный центр информационно-образовательных ресурсов – <http://fcior.edu.ru/>

#### **Электронные библиотечные системы и ресурсы**

9. Электронно-библиотечная система «Консультант студента» – <http://www.studentlibrary.ru/>

10. Руководство по своду знаний по бизнес-анализу (BAВOK 2.0.) (на рус.яз.). URL: <http://iiba.ru/chapter-1-introduction/>

11. Системы дистанционного обучения кафедры экономической кибернетики и прикладной статистики Луганского национального университета имени Владимира Даля в среде Moodle. URL: <https://ecps.gnomio.com/>

#### **Информационный ресурс библиотеки образовательной организации**

12. Научная библиотека имени А. Н. Коняева – <http://biblio.dahluniver.ru/>

#### **8. Материально-техническое и программное обеспечение дисциплины**

Освоение дисциплины ««Методы поиска структурированной и неструктурированной информации»» предполагает использование

академических аудиторий, соответствующих действующим санитарным и противопожарным правилам и нормам.

Прочее: рабочее место преподавателя, оснащенное компьютером с доступом в Интернет.

**Программное обеспечение:**

<b>Функциональное назначение</b>	<b>Бесплатное программное обеспечение</b>	<b>Ссылки</b>
Офисный пакет	OpenOffice 4.3.7	<a href="https://www.openoffice.org/">https://www.openoffice.org/</a>
Операционная система	UBUNTU 19.04	<a href="https://ubuntu.com/">https://ubuntu.com/</a>
Браузер	Firefox Mozilla	<a href="http://www.mozilla.org/ru/firefox/fx">http://www.mozilla.org/ru/firefox/fx</a>
Архиватор	7Zip	<a href="http://www.7-zip.org/">http://www.7-zip.org/</a>
Редактор PDF	Adobe Acrobat Reader	<a href="https://get.adobe.com/ru/reader/">https://get.adobe.com/ru/reader/</a>
Аудиоплеер	VLC	<a href="http://www.videolan.org/vlc/">http://www.videolan.org/vlc/</a>
Графический редактор	GIMP (GNU Image Manipulation Program)	<a href="http://www.gimp.org/">http://www.gimp.org/</a>
Программное обеспечение бизнес-анализа	Microsoft Power BI	<a href="https://powerbi.microsoft.com/ru-ru/downloads/">https://powerbi.microsoft.com/ru-ru/downloads/</a>
Редактор PDF	Adobe Acrobat Reader	<a href="https://get.adobe.com/ru/reader/">https://get.adobe.com/ru/reader/</a>
Аудиоплеер	VLC	<a href="http://www.videolan.org/vlc/">http://www.videolan.org/vlc/</a>
Язык программирования	R	<a href="https://cran.r-project.org/mirrors.html">https://cran.r-project.org/mirrors.html</a>
Среда разработки	RStudio	<a href="https://rstudio.com/products/rstudio/download/">https://rstudio.com/products/rstudio/download/</a>

**9. Оценочные средства по дисциплине**

**Паспорт  
фонда оценочных средств по учебной дисциплине  
«Методы поиска структурированной и неструктурированной  
информации»**

Перечень компетенций (элементов компетенций), формируемых в результате освоения учебной дисциплины (модуля) или практики

<b>№ п/п</b>	<b>Код контролируемой компетенции</b>	<b>Формулировка контролируемой компетенции</b>	<b>Индикаторы достижений компетенции (по реализуемой дисциплине)</b>	<b>Контролируемые темы учебной дисциплины, практики</b>	<b>Этапы формирования (семестр изучения)</b>
1	ПК-5	Способен проводить статистическое наблюдение и	ПК-5.1	Тема 1	7
				Тема 2	
				Тема 3	
				Тема 4	



	группировку, в том числе на основании данных статистических регистров с использованием стандартных методик и технических средств	Тема 5
		Тема 6
		Тема 7
		Тема 8
		Тема 9
		Тема 10
		Тема 11
		Тема 12
		Тема 13
		Тема 14
		Тема 15
		Тема 16
		Тема 17
		Тема 18
		Тема 19
		Тема 20

### Показатели и критерии оценивания компетенций, описание шкал оценивания

№ п/п	Код контролируемой компетенции	Индикаторы достижений компетенции (по реализуемой дисциплине)	Перечень планируемых результатов	Контролируемые темы учебной дисциплины	Наименование оценочного средства
1.	ПК-5	ПК-5.1	Знать сущность и значение информации в развитии современного общества, методы и технологии поиска и обработки экономической информации средствами Интернета и офисных приложений; основные источники информации. ключевые принципы работы с ПК, методов сбора и обработки первичной и вторичной информации из различных источников, в том числе сети Интернет; процесс сбора	Тема 1	Вопросы для обсуждения на практических занятиях (устный опрос), контрольные работы (по вариантам), тесты
				Тема 2	
				Тема 3	
				Тема 4	
				Тема 5	
				Тема 6	
				Тема 7	
				Тема 8	
				Тема 9	
				Тема 10	
				Тема 11	
				Тема 12	

			финансовой, экономической, статистической бухгалтерской информации; возможность обработки собранной информации при помощи информационных технологий; основные источники информации при подготовке аналитического отчета и информационного обзора; основных методов решения аналитических и исследовательских задач. Уметь работать с информацией в глобальных компьютерных сетях; применять при решении прикладных финансово-экономических задач современные информационные технологии для поиска и обработки информации в системе глобальных информационных ресурсов; готовить аналитические обзоры, отчеты и презентации на основе найденной информации; Владеть: основными методами, способами средств поиска, получения, хранения и переработки экономической информации.	Тема 13	
				Тема 14	
				Тема 15	
				Тема 16	
				Тема 17	
				Тема 18	
				Тема 19	
				Тема 20	

**Фонды оценочных средств по дисциплине  
«Методы поиска структурированной и неструктурированной  
информации»**

**Вопросы для обсуждения на практических занятиях  
(устный опрос)**

1. Понятие информации.
2. Характеристики дискретных источников информации.
3. Свойства информации. Условная информация. **Формы адекватности информации. Качества информации.**
4. Структурированная и неструктурированная информация.
5. Понятие информационного поиска.
6. Основные категории информационного поиска: документ, слово, термин, запрос, релевантность, полнота и точность.
7. Этапы информационного поиска.
8. Виды информационного поиска.
9. Методы поиска.
10. Эффективность информационного поиска
11. Сколько поисковых систем существует в интернете?
12. Какие алгоритмы поиска вам известны?
13. Какие дополнительные сервисы предлагают поисковые системы?
14. Назовите российские поисковые системы.
15. Назовите характерные сервисы российских поисковых систем.
16. Назовите лидеров среди российских поисковых систем.
17. Что такое расширенный поиск?
18. Какие команды используют для точного поиска?
19. Как правильно формировать запрос?
20. Что такое релевантность?
21. Как произвести оценку релевантности?
22. Что такое охват?
23. Что такое связывание данных?
24. Как воспользоваться услугой переводчика при поиске информации?
25. Основные стандарты метаданных.
26. Поиск документов различных форматов
27. Информационные потребности научного сообщества. Оценка эффективности поиска. Алгоритмы поиска «по аналогии»
28. Как проводится прямой поиск и поиск по индексу?
29. Обработка булевых запросов.
30. Сравнение расширенной булевой модели и ранжированного поиска

### Критерии и шкала оценивания по оценочному средству «устный опрос»

Шкала оценивания (интервал баллов)	Критерий оценивания
5	Ответ представлен на высоком уровне (студент в полном объеме осветил рассматриваемую проблематику, привел аргументы в пользу своих суждений, владеет профильным понятийным (категориальным) аппаратом и т.п.)
4	Ответ представлен на среднем уровне (студент в целом осветил рассматриваемую проблематику, привел аргументы в пользу своих суждений, допустив некоторые неточности и т.п.)
3	Ответ представлен на низком уровне (студент допустил существенные неточности, изложил материал с ошибками, не владеет в достаточной степени профильным категориальным аппаратом и т.п.)
2	Ответ представлен на неудовлетворительном уровне или не представлен (студент не готов, не выполнил задание и т.п.)

### Контрольные работы (по вариантам)

Задание 1. Определить широту охвата и релевантность по запросам (не менее 5 запросов по 10 поисковым системам), занести данные в таблицу (согласно примерам), провести оценку релевантности.

Задание 2. Произвести приемы простого поиска информации, использование знаков + и -, применение джокера, контекстного поиска (для поисковых систем, где такие команды доступны). При выполнении приемов простого поиска информации показать роль прописных букв, поиск по заголовкам, поиск web-узлов, поиск URL-адресов, поиск ссылок. Осуществить поиск средствами расширенного поиска: OR, AND, NOT, NEAR, вложением команд.

Задание 3.

Необходимо составить списки стоп-слов на основе статистики терминов в коллекции по трем разным коллекциям документов. Стоп-словами будут считаться 5% самых частотных и 5% наименее частотных (по документной частоте,  $\text{document frequency}=\text{df}$ ) терминов из индекса. Предобработка коллекции должна включать основные операции, такие как перевод в нижний регистр и стемминг.

Задание 4.

Для данной коллекции построить графики распределений, иллюстрирующие законы Ципфа и Хипса. Графики должны содержать кривые, построенные по коллекции, и прямые, параметризуемые этими законами и построенными с помощью метода наименьших квадратов.

Задание 5.

Реализовать сжатие словаря фронтальным кодированием. В реализуемой структуре данных для каждого термина должны храниться значения документной частоты, указатели на списки словопозиций, указатели терминов. Необходимо реализовать поиск термина (term lookup):

по данному термину найти его документную частоту. Апробировать на тестовой коллекции.

Задание 6.

Реализовать сжатие файла словопозиций кодированием переменной длины (variable byte encoding). Апробировать на тестовой коллекции.

Задание 7.

Реализовать вариант функции ранжирования tf-idf - модель с опорной нормировкой (pivoted document length normalization).

$$Score(D, Q) = \sum_{t \in D \cap Q} \frac{1 + \ln(tf_{t,D})}{(1 - s) + s \frac{|D|}{avdl}} \cdot tf_{t,Q} \cdot \ln\left(\frac{N + 1}{df_t + 0.5}\right)$$

где D - документ, Q - запрос,  $tf(t,D)$  - частота термина t в документе D, |D| - длина документа D, avdl - средняя длина документа,  $tf(t,Q)$  - частота термина t в запросе, N - количество документов в коллекции,  $df(t)$  - документная частота термина t, s - параметр в диапазоне [0,1]. Апробировать на тестовой коллекции.

Задание 8.

Выберите все опции из ..., при которых HDFS останется в рабочем состоянии, и сохранится доступ ко всем данным. Описание кластера: "Кластер HDFS состоит из 1 NameNode, 1 Secondary NameNode и 6 DataNode. Фактор репликации равняется 3"

Задание 9.

Какие Вы видите проблемы в имплементации метода (transformation) семплирования (sample)? (приведен пример кода на PySpark)

Задание 10.

Может ли HDFS работать если упала NameNode?

Задание 11.

Решить задачу Top100 (WordCount) с помощью MapReduce

Задание 12.

В рамках какой сущности Kafka хранит упорядоченный во времени поток событий?

Задание 13.

Мы работаем с базой данных Cassandra. В настройках указано: фактор репликации - 2, уровень консистенции - quorum. Сколько нод должны успешно ("success") обработать, прежде чем ответить на запрос пользователя?

## Критерии и шкала оценивания по оценочному средству «контрольные работы (по вариантам)»

Шкала оценивания (интервал баллов)	Критерий оценивания
5	Задание выполнено на высоком уровне (правильные ответы даны на 90-100% вопросов/задач)
4	Задание выполнено на среднем уровне (правильные ответы даны на 75-89% вопросов/задач)
3	Задание выполнено на низком уровне (правильные ответы даны на 50-74% вопросов/задач)
2	Задание выполнено на неудовлетворительном уровне (правильные ответы даны менее чем на 50%)

## Тесты

### Задания закрытого типа

1. База знаний является компонентом информационной технологии:
  - 1) экспертных систем;
  - 2) иерархических систем;
  - 3) систем обработки данных.
2. База моделей является компонентом информационной технологии:
  - 1) поддержки принятия решений;
  - 2) ответов на поставленные вопросы;
  - 3) моделирования системы.
3. Главная отличительная черта программ, составляющих интегрированный пакет, является:
  - 1) общий интерфейс пользователя;
  - 2) анализ поставленных задач;
  - 3) эффективность использования.
4. Информация – это:
  - 1) сведения об объектах и явлениях окружающей среды, их параметрах, свойствах и состоянии, которые уменьшают имеющуюся о них степень неопределенности или неполноты знаний;
  - 2) организованный социально-экономический и научно-технический процесс;
  - 3) отыскание рациональных решений в любой сфере;
  - 4) процесс, использующий совокупность средств и методов сбора, обработки и передачи данных.
5. Какая модель имеет структуру в виде дерева и выражает вертикальные связи подчинения нижнего уровня высшему:
  - 1) сетевая;
  - 2) иерархическая;
  - 3) реляционная.
6. Главная цель информатизации:
  - 1) наиболее полное удовлетворение потребностей общества в информации во всех сферах деятельности решать задачи, где известны все ее элементы и взаимосвязи между ними;

2) изменять уровни управления, в зависимости от того, чьи интересы они обслуживают;

7. Данные превращаются в информацию в том случае, если появляется возможность:

- 1) использовать их для уменьшения неопределенности о чем-либо;
- 2) использовать их для увеличения неопределенности о чем-либо;
- 3) использовать их для вычислений.

8. Для автоматизации функций производственного персонала служат ИС:

- 1) управления технологическими процессами (ТП);
- 2) управления производством;
- 3) управления службами технического обеспечения.

9. Для организации и поддержки коммуникационных процессов как внутри организации, так и с внешней средой на базе компьютерных сетей и современных средств работы с информацией служит:

- 1) информационная технология автоматизированного офиса;
- 2) информационная технология обработки данных;
- 3) информационная технология анализа и регулирования;

10. Для решения учетных и финансовых задач используются:

- 1) табличные процессоры;
- 2) текстовые редакторы;
- 3) системы управления базами данных.

11. Для решения хорошо структурированных задач, по которым имеются необходимые входные данные и известны алгоритмы и другие стандартные процедуры их обработки предназначена:

- 1) информационная технология обработки данных;
- 2) информационная технология анализа и регулирования;
- 3) информационная технология автоматизированного офиса.

12. Содержит ли какую-либо информацию таблица, в которой нет ни одной записи?

- 1) пустая таблица не содержит никакой информации;
- 2) пустая таблица содержит информацию о структуре базы данных;
- 3) пустая таблица содержит информацию о будущих записях;
- 4) таблица без записей существовать не может.

13. Реляционная модель требует, чтобы типы данным были...

- 1) простыми;
- 2) структурированными;
- 3) ссылочными.

14. Какие средства управления транзакциями вы знаете? Выберите все ВЕРНЫЕ утверждения

- 1) REVOKE;
- 2) COMMIT;
- 3) ROLLBACK;
- 4) SAVEPOINT;
- 5) EXECUTE.

15. Связывание (link) файла с базой данных означает:

- 1) Установление связи между этим файлом и базой данных Access
- 2) Копирование данных исходного файла в таблицу Access

3) Параллельная обработка данных еще одной программой без преобразования данных в формат Access

16. Принятый способ представления данных: показатели должны быть:

- 1) по строкам;
- 2) по столбцам;
- 3) по ячейкам;
- 4) по диагонали.

17. Интервальные данные – это (подчеркните правильные ответы):

- 1) данные с интервалом;
- 2) данные об интервалах;
- 3) количество измерений в каждом интервале;
- 4) количество интервалов в каждом измерении.

18. Среди ниже приведённых нечисловые данные следующие:

- 1) баллы;
- 2) дихотомические;
- 3) ранги;
- 4) рейтинги.

19. Среди ниже приведённых числовые данные следующие:

- 1) баллы;
- 2) дихотомические;
- 3) ранги;
- 4) рейтинги.

20. Простейшие статистические характеристики – это:

- 1) среднее;
- 2) математическое ожидание;
- 3) с.к.о.;
- 4) дисперсия.

21. Приведение к нормальной форме - это:

- 1) деление на с.к.о.;
- 2) округление;
- 3) деление на среднее;
- 4) деление на константу интегрирования.

22. Какие функции Excel имеют отношение к оцифровке:

- 1) РАНГ;
- 2) КОРРЕЛ;
- 3) СЧЁТЕСЛИ;
- 4) СУММЕСЛИ.

23. Многомерность в статистике - это:

- 1) переменных больше одной;
- 2) переменных больше двух;
- 3) измерений больше 10;
- 4) измерений больше 5.

24. Следующие программы являются специализированными статистическими пакетами:

- 1) EXCEL;
- 2) SPSS;
- 3) GRAPHER;



- 4) STATISTICA.
25. Проверка статистической гипотезы включает в себя:
- 1) ранжирование;
  - 2) принятие уровня значимости;
  - 3) вычисление эмпирического значения;
  - 4) вычисление критического значения.
26. Кластерный анализ предназначен для:
- 1) группировки объектов;
  - 2) группировки показателей;
  - 3) ранжирования объектов;
  - 4) ранжирования показателей.
27. Опции кластерного анализа:
- 1) расстояние между группами;
  - 2) расстояние между показателями;
  - 3) расстояние между объектами;
  - 4) расстояние между телами.
28. Кластерный анализ реализован в программах:
- 1) EXCEL;
  - 2) AGRAPHER;
  - 3) SPSS;
  - 4) STATISTICA.
29. Снижение размерности это:
- 1) уменьшение числа измерений;
  - 2) уменьшение числа объектов;
  - 3) уменьшение числа показателей;
  - 4) уменьшение числа знаков.
30. Компонентный анализ реализован в программах:
- 1) EXCEL;
  - 2) SPSS;
  - 3) AGRAPHER;
  - 4) STATISTICA.
31. Методы, относящиеся к снижению размерности:
- 1) Факторный анализ;
  - 2) компонентный анализ;
  - 3) регрессия;
  - 4) корреляция.
32. Компонентный анализ позволяет:
- 1) сортировать;
  - 2) группировать;
  - 3) ранжировать;
  - 4) упорядочивать.
33. Дихотомическая шкала это:
- 1) состоящая из “да” и “нет”;
  - 2) состоящая из “истина” и “ложь”;
  - 3) состоящая из двух чисел;
  - 4) состоящая из двух рангов.
34. К нечисловым шкалам относятся:

- 1) номинальная;
  - 2) интервалов;
  - 3) абсолютная;
  - 4) ранговая.
35. Существует шкал для описания данных:
- 1) 4;
  - 2) 5;
  - 3) 6;
  - 4) 7.
36. Количество наблюдений - это:
- 1) размерность;
  - 2) объём выборки;
  - 3) ширина;
  - 4) поверхность выборки.
37. Элементы таблицы сопряжённости называются:
- 1) координаты;
  - 2) длины;
  - 3) скорости;
  - 4) частоты.
38. Методы анализа таблиц сопряжённости:
- 1) Критерий Розенбаума;
  - 2) Критерий Колмогорова-Смирнова;
  - 3) хи-квадрат;
  - 4) критерий Фишера.
39. В ходе анализа таблицы сопряжённости выполняется:
- 1) проверка на соответствие;
  - 2) проверка на монотонность;
  - 3) проверка на непротиворечивость;
  - 4) проверка на значимость.
40. Максимальная размерность таблицы сопряжённости может быть:
- 1) 3;
  - 2) 10;
  - 3) 5;
  - 4) какая угодно.
41. Вычисляемое значение критерия хи-квадрат называется:
- 1) Численное значение;
  - 2) экспериментальное значение;
  - 3) реальное значение;
  - 4) эмпирическое значение.
42. Вычисляемое значение хи-квадрат сравнивается с:
- 1) критическим значением;
  - 2) эталонным значением;
  - 3) предельным значением;
  - 4) граничным значением.
43. То, с чем сравнивается вычисляемое значение хи-квадрат, вычисляется в EXCEL функцией:
- 1) ХИ2РАСП;

- 2) ХИ2ОБР;
- 3) ХИ2ТЕСТ;
- 4) ХИ2.
44. К коэффициентам связи относятся:
  - 1) коэффициент контингенции;
  - 2) Коэффициент Чупрова-Крамера;
  - 3) коэффициент ассоциации;
  - 4) коэффициент коллигации.
45. К разновидности критерия хи-квадрат относятся:
  - 1) критерий Вилкоксона;
  - 2) критерий Джонкира;
  - 3) информационный критерий;
  - 4) критерий максимального правдоподобия.
46. Выявление вкладов, вносимых каждой клеткой таблицы, называется:
  - 1) разбиение хи-квадрат;
  - 2) анализ хи-квадрат;
  - 3) локализация хи-квадрат;
  - 4) сортировка хи-квадрат.
47. Лог-линейный анализ – это:
  - 1) анализ синтеза таблиц;
  - 2) статистический анализ связи таблиц;
  - 3) анализ достоверности таблиц;
  - 4) анализ разброса таблиц.
48. Суммарная оперативная память IBM Watson составляет порядка:
  - 1) 100 гигабайт
  - 2) 5000 терабайт
  - 3) 10 зетабайт
  - 4) 15 терабайт
49. Кто ввел термин Большие данные?
  - 1) Клиффорд Линч
  - 2) Алан Тьюринг
  - 3) Бьерн Страуструп
  - 4) Дональд Кнут
50. Какие данные занимают больше мировой памяти относительно остальных?
  - 1) Structured Data
  - 2) Unstructured Data
  - 3) Semi-Structured Data
  - 4) Quasi-Structured Data
51. BigData – это ...
  - 1) Представление фактов, понятий или инструкций в форме, приемлемой для интерпретации, или обработки.
  - 2) Комплексный набор методов обработки структурированных и неструктурированных данных колоссальных объемов.
  - 3) Колоссальный объем данных, собранных человечеством.
  - 4) Класс в Java, предназначенный для хранения данных от 100 Гб

52. Какая компания создала технологию MapReduce?
- 1) Google
  - 2) Yahoo
  - 3) EMC
  - 4) Oracle
53. Данные текстовых файлов с определенными паттернами для их обработки (например, XML) являются:
- 1) Структурированными
  - 2) Полуструктурированными
  - 3) Квазиструктурированными
  - 4) Неструктурированными
54. Что означает термин «Big Data» в информационных технологиях?
- 1) Комплексный набор методов для создания файлов большого объема
  - 2) Комплексный набор методов обработки структурированных и неструктурированных данных колоссальных объемов
  - 3) Файлы с большим количеством данных
  - 4) Представление времени, дня, месяца и года в качестве значения количества миллисекунд, прошедших с начала нашей эры.
55. Данные имеющие определенный тип, формат и структуру (например, транзакционные данные) являются:
- 1) Структурированными
  - 2) Полуструктурированными
  - 3) Квазиструктурированными
  - 4) Неструктурированными
56. Чему примерно равен объем всей существующей на земле информации (в байтах)?
- 1)  $10^{11}$
  - 2)  $10^{21}$
  - 3)  $10^{1010101}$
  - 4)  $10^{171}$
57. В каком году впервые был введен термин Большие данные?
- 1) 2002
  - 2) 2004
  - 3) 2006
  - 4) 2008
58. Что является средством анализа в BI?
- 1) Карты показателей;
  - 2) Совместная работа и управление рабочими процессами;
  - 3) Информационные панели;
  - 4) BI инфраструктура.
59. Основное умение исследователя данных?
- 1) Умение находить наиболее важные элементы в хранимой информации
  - 2) Уметь прогнозировать исход работы системы
  - 3) Находить скрытые логические связи в системе собранной информации
  - 4) Отличать неструктурированные данные от структурированных

60. Какой язык программирования из перечисленных является наиболее важным для аналитика?

- 1) C++
- 2) PHP
- 3) F#
- 4) R

61. Что означает термин «Business Intelligence» в информационных технологиях?

- 1) Комплексный набор методов для создания бизнес планов.
- 2) Методы и инструменты для перевода необработанной информации в осмысленную, удобную для восприятия форму.
- 3) Файлы, содержащие информацию о бизнес-плане.
- 4) Технологии, направленные на развитие бизнеса.

62. Языком, на котором был разработан RabbitMQ, является:

- 1) Java
- 2) Python
- 3) C++
- 4) Erlang

63. Что является главным результатом процесса Business Intelligence?

- 1) Возможность принятия решений для бизнеса
- 2) Результаты интеллектуального анализа данных
- 3) Возможность использования искусственного интеллекта
- 4) Получение структуризации данных после выполнения всех шагов процесса

64. Что из перечисленного не является средством анализа?

- 1) Продвинутое визуализация
- 2) Reporting
- 3) Predictive Modelling
- 4) Data Mining

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

- 1) Генератор нерегламентированных запросов
- 2) Совместная работа и управление рабочими процессами
- 3) Предиктивное моделирование и Data Mining
- 4) Карты показателей

66. Процессом создания и выбора модели для предсказания вероятности наступления некоторого события является:

- 1) OLAP
- 2) Data Mining
- 3) Predictive Modelling
- 4) Data Science

67. Что не является целью процесса Business Intelligence?

- 1) Интерпретация большого количества данных;
- 2) Моделирование исходов различных вариантов действий;
- 3) Модификация существующего программного обеспечения;
- 4) Отслеживание результатов решений.

68. Что из этого не является реализацией Hadoop?

- 1) Google MapReduce
- 2) Phoenix
- 3) GreenMint
- 4) Qizmt

69. Какие из перечисленных пунктов являются достоинствами MapReduce?

- 1) Оптимальная производительность
- 2) Эффективное применение в маленьких кластерах с небольшим объемом данных

- 3) Масштабируемость
- 4) Отказоустойчивость

70. Что такое Oozie?

- 1) Распределенный координационный сервис
- 2) Нереляционная распределенная база данных
- 3) Язык управления потоком данных и исполнительная среда для анализа больших объемов данных

- 4) Сервис для записи и планировки заданий Hadoop

71. Сколько уровней имеет лямбда-архитектура?

- 1) 2
- 2) 3
- 3) 4
- 4) 5

72. Какие компоненты являются частями MapReduce?

- 1) Task Tracker
- 2) Name Node и Data Node
- 3) Job Tracker и Task Tracker
- 4) Job Tracker, Task Tracker, Name Node и Data Node

73. Что такое Spark?

- 1) Инструмент для кластерных вычислений
- 2) Графический движок
- 3) Библиотека для работы с графами
- 4) Технология распределенных вычислений

74. Дайте определение Map Reduce...

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

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

- 3) Распределенная файловая система, работающая на больших кластерах типовых машин

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

75. Что из этого является недостатком MapReduce?

- 1) Фиксированный алгоритм обработки данных
- 2) Масштабируемость
- 3) Отказоустойчивость
- 4) Возможность автоматического распараллеливания

76. Какое API было добавлено в Hadoop v2.0?

- 1) YAWN
- 2) YARN
- 3) SARN
- 4) DARN

77. Какая цель у NameNode в HDFS?

- 1) Хранить индекс того, какая часть данных находится в каком узле
- 2) Хранить имя файла, хранящегося в конкретном узле
- 3) Хранить индекс узла, в котором хранится имя файла
- 4) Хранить имена узлов

78. Вертикальное масштабирование...

- 1) Требуется изменений в прикладных программах, работающих на таких системах
- 2) Не требует никаких изменений в прикладных программах, работающих на таких системах
- 3) Уменьшает производительность каждого компонента БД
- 4) Увеличивает скорость загрузки данных

79. Для достижения какого свойства в БД типа NoSQL нет JOIN операций?

- 1) Intercepting
- 2) Concurrency
- 3) Consistency
- 4) Capacity

80. Что, согласно теореме CAP (теореме Брюера), возможно обеспечить в любой реализации распределённых вычислений?

- 1) Только согласованность данных
- 2) Только доступность данных
- 3) Согласованность данных, доступность данных, устойчивость к разделению
- 4) Не более двух свойств из трёх вышеприведённых

81. Выберите верное определение понятия AP-система:

- 1) Система, во всех узлах которой данные согласованы и обеспечена доступность, жертвует устойчивостью к распаду на секции
- 2) Распределённая система, в каждый момент обеспечивающая целостный результат и способная функционировать в условиях распада
- 3) Распределённая система, отказывающаяся от целостности результата
- 4) Система, автоматически изменяющая данные алгоритма своего с целью сохранения оптимального состояния

82. Что означает термин NoSQL?

- 1) Не SQL
- 2) Не только SQL
- 3) Без SQL
- 4) SQL – плохо

83. 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, это:

- 1) Горизонтальное масштабирование
  - 2) Вертикальное масштабирование
  - 3) Master- slave репликация
  - 4) Peer-to-peer репликация
84. Что из перечисленного относится к графо-ориентированным хранилищам (Graph Store)?
- 1) Neo4j
  - 2) BaseX
  - 3) Elasticsearch
  - 4) Ничего
85. Что поддерживает NoSQL?
- 1) Операцию Insert
  - 2) Полностью стандарт SQL
  - 3) Операцию Join
  - 4) Операцию Group by
86. Какие три свойства фигурируют в определении теоремы CAP?
- 1) Согласованность данных
  - 2) Сложность
  - 3) Доступность
  - 4) Устойчивость к разделению
87. Выделение таблицы или группы таблиц на отдельный сервер это...
- 1) Горизонтальное масштабирование
  - 2) Вертикальное масштабирование
  - 3) Горизонтальный шардинг
  - 4) Вертикальный шардинг
88. Какая из БД на 100% совместима с интерфейсом языка R?
- 1) MySQL R
  - 2) Oracle R
  - 3) PostgreSQL R
  - 4) NoSQL R
89. Что из этого не является типом визуализации?
- 1) График
  - 2) Текст
  - 3) Круговая диаграмма
  - 4) Гистограмма
90. Отображение зависимости значений одной величины от другой - это...
- 1) Матрица
  - 2) График
  - 3) Диаграмма
  - 4) Карта
91. Функция округления до единиц вверх в языке «R»:
- 1) Ceiling(x)
  - 2) Floor(x)
  - 3) Trunc(x)
  - 4) Round(x,2)
92. Что такое сингулярность?



- 1) Точка, в которой функция равна нулю
  - 2) Точка, в которой первая производная равна нулю
  - 3) Точка, в которой вторая производная равна нулю
  - 4) Точка, в которой математическая функция стремится к бесконечности или имеет какие-либо иные нерегулярности поведения
93. Какой тип лицензии у языка R?
- 1) Adware
  - 2) Commercial CC
  - 3) Open source
  - 4) Shareware
94. Какие достоинства у Amazon S3?
- 1) Будет работать всегда
  - 2) Нужно самостоятельно решать сложные задачи распределения файлов между серверами
  - 3) Внезапные всплески популярности не приведут к отказу железа
  - 4) Все вышеперечисленное
95. Что из перечисленного помогает следить за эволюцией документа, над созданием которого работает одновременно большое количество авторов?
- 1) Пространственный поток
  - 2) Исторический поток
  - 3) Визуальный поток
  - 4) Интерактивный поток
96. Преподнесение какой-либо полезной информации в форме интересного рассказа – это...
- 1) Сторителлинг
  - 2) Инфографика
  - 3) Бизнес аналитика
  - 4) Картограмма
97. Что хорошо подходит для дедупликации?
- 1) Картинки, видео, музыка
  - 2) Виртуальные машины
  - 3) Сжатые данные
  - 4) Резервные копии
98. Что является результатом решения задачи регрессии?
- 1) множество допустимых ответов конечно и их называют метками классов
  - 2) допустимым ответом является действительное число или числовой вектор
  - 3) множество допустимых ответов бесконечно
  - 4) алгоритм, принимающий на входе описание объекта
99. Основная цель статистического анализа:
- 1) Поиск генеральной совокупности
  - 2) Выяснение свойств генеральной совокупности
  - 3) Сравнение генеральных совокупностей
  - 4) Выявление последовательности входного набора

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

- 1) Статистический критерий
- 2) Статистическая выборка
- 3) Статистическая гипотеза
- 4) Задача кластеризации

101. К каким алгоритмам классификации относится метод ближайших соседей?

- 1) Метрическим
- 2) Логическим
- 3) Линейным
- 4) Нет верного ответа

102. Преимуществом метода ближайшего соседа является:

- 1) Устойчивость к погрешностям
- 2) Наличие настраиваемых параметров
- 3) Высокое качество классификации
- 4) Простота реализации

103. С помощью какого алгоритма можно найти ассоциативное правило?

- 1) Алгоритм *a priori*
- 2) Алгоритм *k-means*
- 3) Алгоритм *c-means*
- 4) Иерархический алгоритм

104. Технология машинного обучения, когда нет ответов и требуется искать зависимости между объектами, называется ...

- 1) Самостоятельное обучение
- 2) Обучение без учителя
- 3) Обучение с учителем
- 4) Обучение по зависимостям

105. Критерий Пирсона является:

- 1) Критерием значимости
- 2) Параметрическим критерием
- 3) Критерием согласия
- 4) Непараметрическим критерием

106. Чем отличаются ошибки первого и второго рода при принятии решений?

- 1) Ошибка первого рода значительнее, нежели второго
- 2) Ошибка второго рода не обнаруживает различия, которые есть, а первого обнаруживает, которых нет
- 3) Ошибка второго рода значительнее, нежели первого
- 4) Ошибка первого рода не обнаруживает различия, которых нет, а второго обнаруживает

107. Графическая характеристика качества бинарного классификатора ROC-кривая показывает зависимость...

- 1) Величины TPR (доля верных положительных классификаций) от величины FPR (доля ложных положительных классификаций)
- 2) Величины FPR (доля ложных положительных классификаций)

- 3) от величины TPR (доля верных положительных классификаций)  
 4) Величины TNR (доля верных отрицательных классификаций) от величины FPR (доля ложных положительных классификаций)

#### Критерии и шкала оценивания по оценочному средству «тесты»

Шкала оценивания (интервал баллов)	Критерий оценивания
5	Тесты выполнены на высоком уровне (правильные ответы даны на 90-100% тестов)
4	Тесты выполнены на среднем уровне (правильные ответы даны на 75-89% тестов)
3	Тесты выполнены на низком уровне (правильные ответы даны на 50-74% тестов)
2	Тесты выполнены на неудовлетворительном уровне (правильные ответы даны менее чем на 50% тестов)

### Оценочные средства для промежуточной аттестации (зачет с оценкой)

#### Теоретические вопросы

1. Схематизация документа и декодирование последовательности символов.
2. Определение лексикона терминов.
3. Быстрое пересечение инвертированных списков с помощью указателей пропусков.
4. Словопозиции с координатами и фразовые запросы.
5. Лемматизация, морфологический анализ.
6. Принципы работы морфологического анализатора.
7. Процедурный, табличный и вероятностный подходы. Примеры библиотек.
8. Параметрические и зонные индексы.
9. Частота термина и взвешивание.
10. Модель векторного пространства для ранжирования.
11. Варианты функций TF\*IDF.
12. Эффективное ранжирование.
13. Компоненты информационно-поисковой системы.
14. Влияние операторов языка запросов на ранжирование в векторном пространстве.
15. В чем принципиальное отличие концепции Big Data от традиционного подхода BI?
16. Понятие явной (выраженной) и скрытой (структурной) информации.
17. Определение контент-анализа.
18. Каковы основные понятия контент-анализа?
19. Какие существуют виды контент-анализа?
20. Какие существуют этапы контент-анализа?
21. Каковы основные признаки, характеризующие «Большие данные»?
22. Сущность и задачи кластеризации.

23. Основные понятия, принципы и предпосылки генетических алгоритмов.
24. Достоинства и недостатки генетических алгоритмов.
25. Классификация нейронных сетей и принципы построения.
26. Искусственная нейронная сеть прямого прохода.
27. Использование генетических алгоритмов для обучения искусственных нейронных сетей
28. Кластеризация как инструмент предварительной обработки данных для искусственной нейронной сети
29. Какова цель синтаксического анализа?
30. Общая схема алгоритма синтаксического анализа «сверху-вниз» и «снизу-вверх».
31. Схема работы фаз  $map(f, c)$  и  $reduce(f, c)$ .
32. Преимущества, ограничения и недостатки парадигмы MapReduce.
33. Какие бывают модели данных и запросов в NoSQL?
34. Какие бывают системы хранения данных в NoSQL?
35. Основные принципы работы фреймворка Hadoop.
36. Репликация данных в распределенной файловой системе HDFS.
37. Модели развертывания облачных хранилищ.
38. Модели обслуживания облачных хранилищ.
39. Постановка и описание проблемы «последней мили».
40. Безопасность, производительность и надежность при работе с облачными данными.
41. Экономическая составляющая облачных подходов.
42. Способы машинного обучения.
43. Основные фазы обработки «больших данных».
44. Чем отличаются текстовая и персональная базы данных?
45. Метод анализа комбинации слов (collocate analysis).
46. Понятие «сила связи».
47. Статистическая мера совместной встречаемости слов и категорий (Z-score).
48. Реализация закономерностей в системе IBM Cognos Analytics.

Критерии и шкала оценивания по оценочному средству промежуточный контроль («зачет с оценкой»)

Шкала оценивания (интервал баллов)	Критерий оценивания
отлично (5)	Студент глубоко и в полном объёме владеет программным материалом. Грамотно, исчерпывающе и логично его излагает в устной или письменной форме. При этом знает рекомендованную литературу, проявляет творческий подход в ответах на вопросы и правильно обосновывает принятые решения, хорошо владеет умениями и навыками при выполнении практических задач.
хорошо (4)	Студент знает программный материал, грамотно и по сути излагает его в устной или письменной форме, допуская незначительные неточности в утверждениях, трактовках, определениях и категориях или незначительное количество ошибок. При этом владеет необходимыми умениями и навыками при выполнении практических задач.
удовлетворительно (3)	Студент знает только основной программный материал, допускает неточности, недостаточно чёткие формулировки, непоследовательность в ответах, излагаемых в устной или письменной форме. При этом недостаточно владеет умениями и навыками при выполнении практических задач. Допускает до 30% ошибок в излагаемых ответах.
неудовлетворительно (2)	Студент не знает значительной части программного материала. При этом допускает принципиальные ошибки в доказательствах, в трактовке понятий и категорий, проявляет низкую культуру знаний, не владеет основными умениями и навыками при выполнении практических задач. Студент отказывается от ответов на дополнительные вопросы

### Лист изменений и дополнений

№ п/п	Виды дополнений и изменений	Дата и номер протокола заседания кафедры (кафедр), на котором были рассмотрены и одобрены изменения и дополнения	Подпись (с расшифровкой) заведующего кафедрой (заведующих кафедрами)