Рабочая программа дисциплины (модуля)*

Современные технологии поиска и обработки информации

Направление подготовки (специальность)

03.03.02 Физика

Направленность (профиль)

Физика

Присваиваемая квалификация (степень)

бакалавр

Форма обучения

очная

Год(ы) набора 2025

*Рабочая программа дисциплины (модуля) адаптирована для инклюзивного обучения инвалидов и лиц с ограниченными возможностями здоровья

Челябинск 2025 г.

Содержание

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Цель изучения дисциплины - дать студентам теоретические знания и навыки их применения в области поиска информации. Разобрать модели поиска информации, выполнение эффективной индексации текста. Также рассмотрение вопроса о кластеризации и классификации документов.

Изучение дисциплины направлено на развитие следующих индикаторов УК-1.1: " Выполняет поиск информации, определяет критерии системного анализа поставленных задач", ОПК-3.1: "Имеет представление об основных существующих информационных технологиях, используемых при решении профессиональных задач."

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП

Цикл (раздел) ОПОП:	К.М.01.01

2.1 Требования к предварительной подготовке обучающегося:

Дисциплина базируется на школьной программе.

2.2 Дисциплины и практики, для которых освоение данной дисциплины (модуля) необходимо как предшествующее:

Данная дисциплина является основой для дальнейших практик.

Научно-исследовательская работа

Научно-исследовательская работа (получение первичных навыков научно-исследовательской работы)

Преддипломная практика

Подготовка к процедуре защиты и защита выпускной квалификационной работы

3. КОМПЕТЕНЦИИ ОБУЧАЮЩЕГОСЯ, ФОРМИРУЕМЫЕ В РЕЗУЛЬТАТЕ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)

УК-1: Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач

Знать:

Для достижения УК-1.1:
Знать основы выполнения эффективного поиска информации.

Уметь:

Для достижения УК-1.1:
Уметь определять критерии системного анализа для поставленных задач.

Владеть:

Для достижения УК-1.1:
Владеть навыками системного анализа и поиска информации.

ОПК-3: Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности

Знать:

Для достижения ОПК-3.1: Основные существующие информационные технологии, которые используются при решении задач профессиональной деятельности.

Уметь:

Для достижения ОПК-3.1: Использовать существующие информационные технологии для решения задач профессиональной деятельности.

Владеть:

Для достижения ОПК-3.1: Основными существующими информационными технологиями при решении задач профессиональной деятельности.

В результате освоения дисциплины обучающийся должен

3.1	**Знать:**
3.1.1	Основные существующие алгоритмы поиска информации
3.2	**Уметь:**

3.2.1	Пользоваться различными моделями поиска информации
3.3	**Владеть:**
3.3.1	Кластеризации и классификации документов

4. ОБЪЕМ ДИСЦИПЛИНЫ (МОДУЛЯ)

Общая трудоемкость	**2 ЗЕТ**
Часов по учебному плану : 72 в том числе : аудиторные занятия : 36 самостоятельная работа : 32,3 контактная работа: 39,7 ИКР: 3,7	Виды контроля в семестрах: зачеты 1

5. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Код занятия	Наименование разделов и тем /вид занятия/	Семестр / Курс	Часов	Литература
	Раздел 1. Информационный поиск			
1.1	Понятие информационного поиска, его задачи и классификация. /Лек/	1	2	Л2.1 Э1 Э2
1.2	Информационный поиск в вебе: его становление и развитие. Понятие информационного поиска. /Ср/	1	3,3	Л1.1 Э1 Э2
1.3	Оценка информационного поиска. Оценка ранжированных и неранжированных результатов поиска. Оценка релевантности. Оценка информационно-поисковой системы. /Лек/	1	2	Э1 Э2
	Раздел 2. Булев поиск			
2.1	Пример информационного поиска. Обработка булевых запросов. Сравнение расширенной булевой модели и ранжированного поиска.Модель булева поиска. Его особенности и применение. /Лек/	1	4	Л2.1 Э1
2.2	Первая попытка создать инвертированный индекс. /Ср/	1	4	Э2
2.3	Практическое занятие по теме "Булев поиск". Составление матрицы инцидентности по набору документов. Инвертированный список. /Пр/	1	2	Э1 Э2 Э3
	Раздел 3. Инвертированный индекс			
3.1	Понятие и применение инвертированного индекса при поиске информации. /Лек/	1	2	Э1
3.2	Построение индекса. Основы аппаратного обеспечения. Блочное индексирование, основанное на сортировке. /Лек/	1	2	Э2
3.3	Однопроходное индексирование в оперативной памяти. Распределенное и динамическое индексирование, другие типы индексов. /Ср/	1	4	Э1 Э2
	Раздел 4. Электронные библиотечные системы			
4.1	Регистрация в электронных библиотечных системах "Лань" и "Университетская библиотека онлайн". Их возможности, поиск научной литературы в ЭБС. Научная электронная библиотека Elibrary: регистрация и поиск научных статей. /Пр/	1	2	Э1 Э2 Э3
	Раздел 5. Поисковые системы			
5.1	Особенности различных поисковых систем, принципы их работы. /Пр/	1	2	Э1 Э2 Э3
5.2	Изучение принципов и алгоритмов, на которые опираются поисковые системы. /Ср/	1	4	Э1 Э2 Э3
	Раздел 6. Основы поиска в вебе			

6.1	Характеристики веба,опыт пользователей поисковых систем. /Ср/	1	3	Э1 Э2
	Раздел 7. Лексикон и списки словопозиций			
7.1	Схематизация документа и декодирование последовательности символов. Определение лексикона терминов. /Лек/	1	2	Э1 Э2
7.2	Быстрое пересечение инвертированных списков с помощью указателей пропусков. /Ср/	1	2	Э1 Э2
7.3	Практические занятия по темам "Лексикон и списки словопозиций" и "Фразовые запросы". Поэтапное составление инвертированного индекса, координатный индекс. /Пр/	1	4	Э1 Э2 Э3
	Раздел 8. Словари и нечеткий поиск			
8.1	Поисковые структуры для словарей. Запросы с джокером. Исправление опечаток. Фонетические исправления. /Лек/	1	4	Э1 Э2
8.2	Методы обработки запросов, содержащих орфографические ошибки и другие неточности. /Ср/	1	6	Э1 Э2
8.3	Практические занятия по темам "Поиск по сходству" и "Фонетические исправления". Обработка запросов, содержащих орфографические ошибки. /Пр/	1	4	Э1 Э2
	Раздел 9. Сжатие индекса			
9.1	Сжатие инвертированного индекса. Алгоритмы сжатия словаря и списка словопозиций. /Пр/	1	4	Э1
9.2	Алгоритмы сжатия инвертированного индекса. /Ср/	1	6	Э2 Э3
	Раздел 10. Иная контактная работа			
10.1	Индивидуальные консультации, текущий контроль /ИКР/	1	3,7	Л1.1Л2.1 Э1 Э2 Э3

6. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

6.1. Перечень видов оценочных средств

Самостоятельная работа №1;
Самостоятельная работа №2;
Самостоятельная работа №3;
Самостоятельная работа №4;
Самостоятельная работа №5;
Самостоятельная работа №6.

6.2. Типовые контрольные задания и иные материалы для текущей аттестации

Типовые задания для самостоятельной работы:
см. приложение

6.3. Типовые контрольные вопросы и задания для промежуточной аттестации

1. Понятие информационного поиска, его цель и задачи;
2. Классификация информационного поиска;
3. Модель булева поиска;
4. Определение эффективности системы информационного поиска: точность и полнота;
5. Инвертированный индекс. Этапы его построения;
6. Оптимизация запроса;
7. Структурная единица документа. Проблема детализации индексирования;
8. Лексема и термин: определение и различия;
9. Стоп-слова: определение, способ создания списка стоп-слов, использование стоп-слов в системах информационного поиска;
10. Нормализация лексем. Классы эквивалентности;
11. Индексация ударений и диакритических знаков;
12. Индексирование заглавных букв;
13. Стемминг. Пример;
14. Лемматизация. Пример;

15. Указатели пропуска;
16. Фразовые запросы. Двухсловные индексы;
17. Фразовые запросы. Координатный индекс;
18. Комбинированная схема обработки фразовых запросов;
19. Реализация словаря. Хеширование;
20. Реализация словаря. Деревья поиска;
21. Запросы с джокером и их обработка;
22. K-граммный индекс для обработки запросов с джокером;
23. Реализация исправления опечаток;
24. Фонетические исправления. Soundex-индекс;
25. Характеристики аппаратного обеспечения, влияющие на обработку запросов;
26. Архитектура MapReduce;
27. Сжатие словаря: цель и методы;
28. Сжатие инвертированного файла;
29. Классификация текстов.

6.4. Критерии оценивания

Зачет проводится в присутствии преподавателя и предполагает краткий ответ на вопросы. Вопросы составляются с учётом материала, пройденного на лекционных занятиях. Итоговая оценка выставляется по балльной системе.
Студенту необходимо ответить на два теоретических вопроса. Каждый ответ оценивается в 5 баллов. Для того, чтобы получить максимальное количество баллов, необходимо предоставить полный ответ на вопрос. Суммируются баллы, полученные на зачете (10 максимум), и баллы, полученные на самостоятельные работы. Для получения максимального количества баллов необходимо выполнить все задания самостоятельной работы без ошибок в установленные сроки, ответить на вопросы преподавателя во время защиты работы. Каждая самостоятельная работа оценивается от 0 до 10 баллов. Частичное выполнение заданий, допущенные ошибки при их выполнении или при ответе на вопросы преподавателя приводят к снижению количества баллов за самостоятельную работу. Полученные студентами баллы суммируются, итоговая оценка выставляется исходя из полученной суммы баллов:
От 0 до 40 баллов – «незачтено»
От 41 до 50 баллов – «зачтено».

7. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

7.1. Рекомендуемая литература

7.1.1. Основная литература

	Авторы, составители	Заглавие	Издательство, год	Ресурс
Л1.1	Громов Ю. Ю., Дидрих И. В., Иванова О. Г., Ивановский М. А., Однолько В. Г.	Информационные технологии: учебник (https://biblioclub.ru/index.php?page=book&id=444641)	Тамбов : Тамбовский государственный технический университет (ТГТУ), 2015	ЭБС

7.1.2. Дополнительная литература

	Авторы, составители	Заглавие	Издательство, год	Ресурс
Л2.1	Гасанов Э. Э., Кудрявцев В. Б.	Теория хранения и поиска информации (https://znanium.com/catalog/document?id=259903)	Москва : Издательская фирма "Физико-математическая литература" (ФИЗМАТЛИТ), 2002	ЭБС

7.2. Перечень ресурсов информационно-телекоммуникационной сети "Интернет"

Э1	Лань [Электронный ресурс] : электронно-библиотечная система (ЭБС) / издательство Лань. – URL: http://e.lanbook.com/.
Э2	Университетская библиотека онлайн [Электронный ресурс] : электронно-библиотечная система (ЭБС) / ООО Директмедиа Паблишинг. – URL: http://biblioclub.ru/.
Э3	eLIBRARY.RU [Электронный ресурс] : электронная библиотека / Науч. электрон. б-ка. – URL: http://elibrary.ru/defaultx.asp.

7.3 Перечень информационных технологий

7.3.1 Программное обеспечение

Adobe Reader

WinDjView

LMS Moodle

LibreOffice

7.3.2 Профессиональные базы данных и информационно-справочные системы

1. Электронный каталог научной библиотеки ЧелГУ [Электронный ресурс] : база данных / Челяб. гос. ун-т. – Челябинск, 1992 .

2. eLIBRARY.RU : научная электронная библиотека : сайт. – Москва, 2000 – . – URL: https://elibrary.ru – Режим доступа: для зарегистрир. пользователей. – Текст : электронный.

3. Mathematical Reviews (MR) : реферативная база данных / American Mathematical Society. – URL: http://www.ams.org/mathscinet/ – Яз. рус., англ. – Режим доступа: для зарегистрир. пользователей ЧелГУ. – Текст : электронный.

8. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Для реализации дисциплины используются учебные аудитории для проведения занятий лекционного типа, занятий семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы.

Учебные аудитории укомплектованы специализированной мебелью и техническими средствами обучения (компьютерная техника с подключением к сети "Интернет" для практических занятий).

Для проведения занятий лекционного типа предлагаются наборы демонстрационного оборудования и учебно- наглядных пособий, такие как презентации.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с подключением к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду университета

9. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)

В результате изучения теоретических основ дисциплины и реализации в процессе обучения требований по прикладной направленности дисциплины, а также самостоятельной работы студент должен знать, уметь и владеть составляющими компетенций, определенных в программе.
Основными видами учебных занятий являются: лекции, практические занятия, самостоятельная работа и зачет.
Аудиторные лекции студента имеют своей целью формирование целостной системы знаний по изучаемому предмету.
Студент может воспользоваться основной и дополнительной литературой.
Самостоятельная работа студента начинается с внимательного ознакомления с программой данной дисциплины.
Требуется творческое отношение к самой Программе учебного курса. Вопросы, составляющие ее содержание, обладают разной степенью важности. Есть вопросы, выполняющие функцию логической связки содержания темы и всего курса, имеются вопросы описательного или разъяснительного характера. Эти вопросы не составляют сути, понятийного, концептуального содержания темы, но необходимы для целостного восприятия изучаемых проблем. Успешно освоив теоретический материал, студент будет готов к проведению практических заданий, которые рассматриваются как дальнейшее углубление и расширений знаний по предмету.
В освоении дисциплины (модуля) инвалидами и лицами с ограниченными возможностями здоровья большое значение имеет индивидуальная работа. Под индивидуальной работой подразумевается две формы взаимодействия с преподавателем: индивидуальная учебная работа (консультации), т.е. дополнительное разъяснение учебного материала и углубленное изучение материала с теми обучающимися, которые в этом заинтересованы, и индивидуальная воспитательная работа.
Индивидуальные консультации по предмету является важным фактором, способствующим индивидуализации обучения и установлению воспитательного контакта между преподавателем и обучающимся инвалидом или обучающимся с ограниченными возможностями здоровья.
В случае применения при обучении дисциплины электронного обучения, дистанционных образовательных технологий общение обучающихся и преподавателя осуществляется в режиме реального времени (чаты, видео- конференции и др.) или отложенного времени (система дистанционного обучения Moodle, форумы, электронная почта).
Большую часть времени обучающиеся самостоятельно работают с учебно-методическими материалами. Студенты имеют возможность консультироваться с преподавателем по всем вопросам, возникающим в ходе самостоятельной работы посредством электронной почты, социальных сетей, Moodle.
Доступ обучающегося к учебным ресурсам в режиме отложенного времени, самостоятельной работы осуществляется через сеть Интернет в удобном для него месте, времени и темпе.

При обучении лиц с ограниченными возможностями здоровья электронное обучение, дистанционные образовательные технологии предусматривают возможность приема-передачи информации в доступных для них формах.

Реализация дисциплины с применением электронного обучения, дистанционных образовательных технологий (далее – ЭО, ДОТ) осуществляется на основании «Положения о реализации основных и дополнительных образовательных программ с применением электронного обучения и дистанционных образовательных технологий в федеральном государственном бюджетном образовательном учреждении высшего образования «Челябинский государственный университет», «Положения о порядке зачета обучающимися по основным профессиональным образовательным программам высшего образования в ФГБОУ ВО «ЧелГУ» результатов освоения в организациях, осуществляющих образовательную деятельность, учебных предметов, курсов, дисциплин (модулей), практик, дополнительных образовательных программ» посредством электронной информационно-образовательной среды ФГБОУ ВО «ЧелГУ». В исключительных случаях (форс-мажор и т.п.) при реализации образовательной деятельности с применением ЭО, ДОТ могут применять компоненты, не входящие в перечень электронной информационно- образовательной среды.

10. СПЕЦИАЛЬНЫЕ УСЛОВИЯ ОСВОЕНИЯ ДИСЦИПЛИНЫ ОБУЧАЮЩИМИСЯ С ИНВАЛИДНОСТЬЮ И ОГРАНИЧЕННЫМИ ВОЗМОЖНОСТЯМИ ЗДОРОВЬЯ

Освоение дисциплины инвалидами и лицами с ограниченными возможностями здоровья осуществляется с использованием специальных технических средств и информационных технологий, предоставляемых Ресурсным учебно-методическим центром по обучению инвалидов и лиц с ограниченными возможностями здоровья ЧелГУ по запросу обучающегося (мобильные специальные технические средства для лиц с нарушениями зрения и с нарушением слуха, ассистивные информационные технологии).
При необходимости для обучающихся с нарушениями зрения на рабочих местах для проведения практических или лабораторных занятий устанавливается специальное программное обеспечение (программа речевой навигации, речевые синтезаторы, экранные лупы).
В учебные аудитории обеспечивается беспрепятственный доступ для обучающихся с инвалидностью и с ограниченными возможностями здоровья. В каждой аудитории, где обучаются инвалиды и лица с ограниченными возможностями здоровья, предусматривается соответствующее количество мест для обучающихся с учетом нарушений их здоровья.
Для освоения дисциплины инвалидам и лицам с ограниченными возможностями здоровья предоставляется доступ к печатным источникам, имеющимся в научной библиотеке ЧелГУ, с помощью специальных технических средств; доступ с помощью специальных технических и программных средств к электронным источникам, представленным в форме электронного документа в фонде научной библиотеки ЧелГУ или электронно-библиотечных системах.
Учебно-методические материалы для обучающихся из числа инвалидов и лиц с ограниченными возможностями здоровья предоставляются в формах, адаптированных к ограничениям их здоровья и особенностям восприятия информации.
Для инвалидов и лиц с ограниченными возможностями здоровья освоение дисциплины может быть частично или полностью осуществлено с использованием дистанционных образовательных технологий.
При проведении промежуточной аттестации по дисциплине обучающимся с инвалидностью и с ограниченными возможностями здоровья обеспечивается по их заявлению предоставление в доступной форме в зависимости от их индивидуальных особенностей инструкции о порядке проведения промежуточной аттестации, оценочных средств и возможности ответов на задания (письменно на бумаге, набор ответов на компьютере, письменно шрифтом Брайля, с использованием услуг ассистента, устно).
При проведении процедуры оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья предусматривается использование предоставленных ЧелГУ или собственных технических средств, необходимых им в связи с их индивидуальными особенностями. При необходимости инвалидам и лицам с ограниченными возможностями здоровья предоставляется дополнительное время для подготовки ответа на задания, процедура оценивания результатов обучения по дисциплине может проводиться в несколько этапов.

Самостоятельная работа №1 по теме «Булев поиск».

Вариант 1

1. Прочитайте следующие документы:

Документ 1.

В Интернете с каждым днём скапливается всё больше информации, когда-либо созданной и вновь создаваемой людьми. Равнодоступность большей части информации в Интернете уравнивает возможности доступа к этой информации как обычных пользователей Интернета и журналистов локальных СМИ, так и сотрудников мировых информационных агентств. Благодаря Интернету перед каждым человеком ежедневно и даже ежесекундно открывается доступ к многомиллионной аудитории, которой он может передать свой информационный материал, полученный, например, с помощью обычного мобильного телефона с диктофоном и встроенной фотокамерой. Следовательно, уровень монополизации деятельности по распространению информации также снижается благодаря Интернету.

Документ 2.

До недавнего времени ограничения в прямой коммуникации между людьми, порождаемые пространством и временем, во многом определяли потребность людей в услугах журналистов. По мере роста общего количества пользователей Интернета, а среди них – числа владеющих английским языком, эти ограничения всё в большей степени снимаются, что закономерно ведёт к уменьшению спроса на услуги журналистов. Одновременно с этим растёт объем «сырой» информации, доступной каждому отдельному пользователю Интернета, что актуализирует проблему её отбора и редактирования. Последнее всегда входило в перечень функций журналистики, но с ростом числа пользователей Интернета эффективный информационный поиск начинает приобретать всё большую значимость не только в журналистской деятельности, но и в других разнообразных сферах общественной деятельности. Таким образом, информационный поиск – это процесс поиска неструктурированной документальной информации.

Документ 3.

Поиск информации представляет собой процесс выявления в некотором множестве документов (текстов), которые посвящены заданной теме (предмету) и удовлетворяют заранее определенному условию поиска (запросу), а также содержат необходимые (соответствующие информационной потребности) факты, сведения и данные. Процесс поиска включает последовательность операций, направленных на сбор, обработку и предоставление необходимой информации заинтересованным лицам.

Документ 4.

Комплекс программ, предназначенных для информационного поиска, называется поисковой машиной. Обычно является частью поисковой системы – автоматизированного программно-аппаратного комплекса с веб-интерфейсом, предоставляющего возможность поиска информации в Интернете. Самая известная поисковая система в мире – это Google, самая популярная в России – Яндекс, а одной из самых старых поисковых систем является Yahoo. Как уже было отмечено ранее, в архитектуре поисковой системы можно выделить поисковую машину – ядро системы, представленное набором программных модулей; базу данных или индекс, хранящую информацию обо всех известных поисковой системе Интернет ресурсах; и

набор сайтов, являющих собой точки входа пользователей в систему. Все это соответствует классической трехуровневой архитектуре информационных систем: есть пользовательский интерфейс, бизнес логика, которая в данном случае представлена реализацией алгоритмов поиска и база данных.

Документ 5.

Для того, чтобы найти в Интернете требуемую информацию, необходимо знать либо адрес её местоположения (например, адрес html-страницы или файла), либо пользователя Интернета, который может предоставить информацию. Если мы не знаем ни адреса, ни человека, который мог бы нам помочь, то следует перейти к вопросам «Как можно узнать адрес размещения информации?» или «Как найти человека, который мог бы нам помочь с поиском информации?». При этом не следует переоценивать возможности Интернета. Лучшие результаты может дать совмещение онлайновых и оффлайновых методов поиска информации.

Документ 6.

Сегодня существует достаточно большое количество методов информационного поиска в Интернете и через Интернет. В каждом конкретном случае успешность поиска определяется знаниями возможных методов и навыками владения ими, знанием этнических языков, на которых эта информация может быть представлена, либо нашими социальными связями.

2. Постройте матрицу инцидентности «термин — документ» для терминов:
Интернет, поиск, запрос, индекс, информация, система, машина. Словоформа и регистр значения не имеют. Термины расположите в таблице в алфавитном порядке.

3. Обработайте запрос *поиск* AND *информация* AND NOT *интернет*, взяв векторы для терминов из матрицы инцидентности и выполнив поразрядные операции AND и NOT. В ответе укажите список документов.

4. Составьте инвертированный список для коллекции документов Вашего варианта используя в качестве словаря термины из пункта 2.

Самостоятельная работа №2
по теме «Лексикон и списки словопозиций»

Вариант 1

1. Разбейте текст каждого документа на лексемы:

Документ 1.

Состав и интерпретация поддерживаемых метасимволов. Часто называется «диалектом» регулярного выражения.

Документ 2.

Особенности взаимодействия регулярных выражений с языком или программой.
Документ 3.

Специфика применения регулярных выражений к тексту.

2. Выполните предварительную лингвистическую обработку лексем с помощью нормализации лексем и игнорирования стоп-слов (союзы, предлоги).

3. Составьте инвертированный индекс для данной коллекции документов с указанием частоты и списка словопозиций для каждого термина. Термины расположите в алфавитном порядке.

4. Нормализуйте лексемы в запросе: *регулярные* AND *выражения* и обработайте его.

Самостоятельная работа №3
по теме «Словопозиции с координатами и фразовые запросы»

Вариант 1

1. Разбейте текст каждого документа на лексемы:

Документ 1.

Как правило, поисковые машины поддерживают два режима: режим простого поиска и режим расширенного поиска.

Документ 2.

Можно просто вводить через пробел одно или несколько слов; поиск слов со всевозможными окончаниями моделируется символом * в конце слова.

Документ 3.

Многие системы позволяют искать словосочетания или фразу, для этого необходимо ее заключить в кавычки.

Документ 4.
Возможно обязательное включение или исключение определенных слов.

Документ 5.

Основная проблема поиска по примитивно составленному запросу заключается в том, что поисковая машина найдет все страницы, на которых указанные слова встречаются в любой части документа.
Документ 6.

Как правило, количество найденных страниц будет слишком велико.

2. Составьте координатный индекс для данной коллекции документов в формате термин док1: <позиция1, позиция2, …>; док2: <позиция1, позиция2, …>; и т.д., используя следующий словарь терминов: поиск, машина, запрос, режим, слово, документ, страница. Термины расположите в алфавитном порядке.

3. Какие документы удовлетворяют фразовому запросу поисковая машина?

4. Укажите набор документов, удовлетворяющих запросу: режим /2 поиск.

Самостоятельная работа №4 по теме «Поиск по сходству»

Вариант 1

1. С помощью перестановочного индекса определите термины лексикона ***поисковик, иск, поиск, постфикс, пост, пуск***, соответствующие запросу ***по*ск***.

2. Используя алгоритм динамического программирования, вычислите расстояние редактирования между строками ***сильный*** и ***стильный***.

3. Составьте биграммный индекс терминов лексикона ***стена, струна, сторона, оборона, рана, охрана*** и найдите термины, содержащие по крайней мере две из всех биграмм в запросе ***страна***.

4. Найдите коэффициент Жаккара для строк запроса и каждого термина, полученного в предыдущем пункте. Для каких терминов этот коэффициент наибольший?

Самостоятельная работа №4 по теме «Фонетические исправления»

Вариант 1

1. Используя алгоритм фонетического хеширования и таблицу кодирования русских букв составьте soundex-индекс для следующего словаря терминов:

собор, луг, забор, сбор, информатика, запор, инфракрасный, инвентарь, зубр, лук

Таблица кодирования русских букв

0. у, е, ё, ы, а, о, э, я, и, ю, ь, ъ
1. б, п
2. ф, в
3. ж, з, с, х
4. к, г
5. ц, ч, ш, щ
6. д, т
7. л, й
8. м, н
9. р

2. Какие термины имеют одинаковые Soundex-коды?

3. Преобразуйте термин запроса **СОБР** в Soundex-код и выполните поиск по soundex-индексу.

Самостоятельная работа №6 по теме «Сжатие индекса»
Вариант 1

1. Создайте блочное хранение (k = 4) для следующего словаря терминов:
поиска, информации, поиску, сбора, информационный, поисковой, сборный, информация, сбору, поисковик, сборного, информацию.

2. Выполните дальнейшее сжатие с помощью фронтальной упаковки. Для обозначения общего префикса используйте символ ◊ (Alt-код символа Alt+9674).

3. Вычислите коды, полученные с помощью схемы байтового кодирования для следующего инвертированного списка: <4, 10, 11, 12, 15, 62, 63, 265, 268, 270, 400> Запишите бинарные коды блоками по 8 бит.

03.03.02 Физика, направленность Физика, РПД по дисциплине «Современные технологии поиска и обработки информации», 2025 год набора, форма обучения очная

Проректор по учебной работе	утверждено 24.02.25	А.А. Саламатов

Ученым советом физического факультета

Протокол заседания № 06 от 20.02.2025

Председатель Ученого совета физического факультета	согласовано	М.А. Загребин

Заседанием кафедры теории управления и оптимизации

Протокол заседания № 8 от 20.02.2025

Заведующий кафедрой	согласовано	И.В. Изместьев
Автор (составитель)		Е.А. Скрипов

Структура рабочей программы соответствует приказу ректора ФГБОУ ВО «ЧелГУ» от «13» апреля 2021 г. № 247-1