



**МИНОБРНАУКИ РОССИИ**  
федеральное государственное бюджетное образовательное учреждение  
высшего образования  
**«ИРКУТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**  
ФГБОУ ВО «ИГУ»  
**Кафедра Алгебраических и информационных систем**

**«УТВЕРЖДАЮ»**  
Директор ИМИТ ИГУ  
  
**М.В. Фалалеев**  
**«19» мая 2021 г.**



### **Рабочая программа дисциплины**

**Наименование дисциплины (модуля)** Б1.В.ДВ.03.02 Алгоритмы для интернет

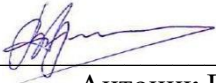
**Направление подготовки** 09.03.03 Прикладная информатика

**Направленность (профиль) подготовки** Проектирование и разработка информационных систем


**Квалификация выпускника** – бакалавр

**Форма обучения** очная

Согласовано с УМК Института математики  
и информационных технологий  
Протокол № 4 от «19» мая 2021 г.

Председатель   
\_\_\_\_\_ Антоник В.Г.

Рекомендовано кафедрой Алгебраических и  
информационных систем ИМИТ ИГУ:  
Протокол № 11 от «29» апреля 2021 г.

Зав. кафедрой   
\_\_\_\_\_ Пантелеев В.И.

## СОДЕРЖАНИЕ

1. Цели и задачи дисциплины	42.	Место дисциплины в структуре опоп во	43.
Требования к результатам освоения дисциплины	44.	Содержание и структура дисциплины	64.1.
СОДЕРЖАНИЕ ДИСЦИПЛИНЫ, СТРУКТУРИРОВАННОЕ ПО ТЕМАМ, С УКАЗАНИЕМ ВИДОВ УЧЕБНЫХ ЗАНЯТИЙ И ОТВЕДЕННОГО НА НИХ КОЛИЧЕСТВА АКАДЕМИЧЕСКИХ ЧАСОВ	64.2.	План внеаудиторной самостоятельной работы обучающихся по дисциплине	64.3.
СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА	74.4.	МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ОРГАНИЗАЦИИ САМОСТОЯТЕЛЬНОЙ РАБОТЫ СТУДЕНТОВ	104.5.
ПРИМЕРНАЯ ТЕМАТИКА КУРСОВЫХ РАБОТ (ПРОЕКТОВ)	125.	Учебно-методическое и информационное обеспечение дисциплины (модуля)	136.
Материально-техническое обеспечение дисциплины	137.	Образовательные технологии	138.
Оценочные материалы для текущего контроля и промежуточной аттестации	13		

## 1. Цели и задачи дисциплины

### Цель

Знакомство с приемами решения практических задач в сети интернет с помощью алгоритмов поиска и индексации данных.

### Задачи:

Изучение моделей и алгоритмов информационного поиска.

## 2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП ВО

2.1. Учебная дисциплина (модуль) относится к части программы, формируемой участниками образовательных отношений, и изучается на третьем курсе.

2.2. Для изучения данной учебной дисциплины (модуля) необходимы знания, умения и навыки, сформированные Информатика, Программирование, Алгоритмы и анализ сложности.

2.3. Перечень последующих учебных дисциплин, для которых необходимы знания, умения и навыки, формируемые данной учебной дисциплиной: не предусмотрены.

## 3. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Процесс освоения дисциплины направлен на формирование компетенций (элементов следующих компетенций) в соответствии с ФГОС ВО по соответствующему направлению подготовки.

### Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с индикаторами достижения компетенций

Компетенция	Индикаторы компетенций	Результаты обучения
ПК-3 Способность демонстрации общенаучных базовых знаний математических и естественных наук и информационных технологий; способность применять в профессиональной деятельности современные языки программирования, операционные системы, электронные библиотеки и пакеты программ, сетевые технологии.	ИДК ПК3.1 Способен использовать математические знания в профессиональной деятельности	Знает модели информационного поиска Умеет формализовать задачу в рамках одной из моделей информационного поиска Владеет навыками описания коллекции документов в рамках информационной модели
	ИДК ПК3.2 Способен использовать теоретические принципы информационных технологий в профессиональной деятельности	Знает принципы построения поисковой системы, методы сжатия поискового индекса, методы оценки качества работы поисковой системы, задачи информационного поиска Умеет выполнять сжатие поискового индекса Владеет навыками описания коллекции документов, построения индекса,

		выполнения запросов в выбранной информационной модели
	ИДК ПК3.3 Способен применять в профессиональной деятельности современные языки программирования, операционные системы пакеты программ и сетевые технологии	Знает механизмы индексации текстовых и гипертекстовых документов на современном языке программирования Умеет реализовывать построение поискового индекса на современном языке программирования Владеет навыками реализации поискового индекса на современном языке программирования

#### 4. СОДЕРЖАНИЕ И СТРУКТУРА ДИСЦИПЛИНЫ

Объем дисциплины составляет 2 зачетных единиц, 72 часа, практическая подготовка 72.

Форма промежуточной аттестации: 5 семестр - зачет.

##### 4.1. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ, СТРУКТУРИРОВАННОЕ ПО ТЕМАМ, С УКАЗАНИЕМ ВИДОВ УЧЕБНЫХ ЗАНЯТИЙ И ОТВЕДЕННОГО НА НИХ КОЛИЧЕСТВА АКАДЕМИЧЕСКИХ ЧАСОВ

№ п/п	Раздел дисциплины/темы	Се мес тр	Виды учебной работы, включая самостоятельную работу обучающихся и трудоемкость (в часах)				Формы текущего контроля успеваемости
			Контактная работа преподавателя с обучающимися			Самостоя тельная работа + контроль	
			Лекции	Семинарск ие (практичес кие занятия)	Контроль обучения		
1	Тема 1. Введение	5	2		1		
2	Тема 2. Булева модель ИП	5	2	2	1	2	лаб.
3	Тема 3. Лексикон и списки словопозиций	5	2	2	1	4	лаб.
4	Тема 4. Оценка ИП	5	2		1		лаб.
5	Тема 5. Сжатие индекса	5	2	2	1	4	лаб.
6	Тема 6. Векторная модель ИП	5	4	4	1	4	лаб.
7	Тема 7. Анализ Ссылок	5	2	2	1	2	лаб.
8	Тема 8. Задачи ИП	5		4	1	16	доклад
<b>Итого часов</b>			16	16	8	32	

##### 4.2. План внеаудиторной самостоятельной работы обучающихся по дисциплине

Семестр	Название раздела, темы	Самостоятельная работа обучающихся			Оценочное средство	Учебно-методическое обеспечение самостоятельной работы
		Вид самостоятельной работы	Сроки выполнения	Затраты времени (час.)		
5	Тема 2. Булева модель ИП	Выполнение практической работы	30.09	2	Проверка домашней работы	Материалы курса на платформе ИОС DOMIC
5	Тема 3. Лексикон и списки словопозиций	Выполнение практической работы	14.10	4	Проверка домашней работы	Материалы курса на платформе ИОС DOMIC
5	Тема 5. Сжатие индекса	Выполнение практической работы	28.10	4	Проверка домашней работы	Материалы курса на платформе ИОС DOMIC
5	Тема 6. Векторная модель ИП	Выполнение практической работы	18.11	4	Проверка домашней работы	Материалы курса на платформе ИОС DOMIC
5	Тема 7. Анализ Ссылок	Выполнение практической работы	9.12	2	Проверка домашней работы	Материалы курса на платформе ИОС DOMIC
5	Тема 8. Задачи ИП	Подготовка доклада	23.12	16	Заслушивание доклада, ответы на вопросы по теме доклада	Материалы курса на платформе ИОС DOMIC
Общая трудоемкость самостоятельной работы по дисциплине (час)				32		
Из них объем самостоятельной работы с использованием электронного обучения и дистанционных образовательных технологий (час)				32		

#### 4.3. СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

#### Тема 1. Введение

Проблемы развития Интернета. Структурированные и неструктурированные данные. Термин информационный поиск. Запрос. Объект запроса. Парадигмы поиска. Этапы информационного поиска. Виды поиска.

#### Тема 2. Булева модель ИП

Создание инвертированного индекса. 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.

#### Тема 4. Оценка ИП

Меры информационно-поисковых систем: полнота, точность, F-мера. Оценка неранжированных результатов поиска. Оценка ранжированных результатов поиска. Стандартные текстовые коллекции. Критерии оценки поисковых систем. Согласованность экспертных оценок.

Статические и динамические аннотации.

#### Тема 5. Сжатие индекса

Статистические характеристики термов в информационном поиске. Сжатие словаря. Сжатие инвертированного файла.

#### Тема 6. Векторная модель ИП

Параметрические и зонные индексы. Частота термина и взвешивание. Модель векторного пространства для ранжирования. Варианты функции tf-idf.

#### Тема 7. Анализ Ссылок

Ориентированный веб-граф. Индексация текста ссылок. Применение анализа цитирования. Модель случайного блуждания. Алгоритм PageRank. Алгоритм HITS.

#### Тема 8. Задачи ИП

Индексация звука на примере Shazam. Обзор открытых технических решений для исправления опечаток в запросах. Рекомендующие системы. Совместная фильтрация

Индексация изображений. Библиотеки для высокоскоростного полнотекстового поиска Lucene и Elasticsearch. Вопросно-ответные системы.

Поиск в полуструктурированных данных. Система полнотекстового поиска Sphinx. Использование речевых технологий в поисковых системах

### 4.3.1. Перечень семинарских, практических занятий и лабораторных работ

№ п/п	№ раздела и темы	Наименование семинаров, практических и лабораторных работ	Трудоемкость (час.)		Оценочные средства	Формируемые компетенции (индикаторы)*
			Всего часов	Из них		



				практическая подготовка		
1	2	3	4	5	6	7
1	2	Тема 2. Булева модель ИП	2	2	Проверка лабораторной работы	ПК-3
2	3	Тема 3. Лексикон и списки словопозиций	2	2	Проверка лабораторной работы	ПК-3
3	5	Тема 5. Сжатие индекса	2	2	Проверка лабораторной работы	ПК-3
4	6	Тема 6. Векторная модель ИП	4	4	Проверка лабораторной работы	ПК-3
5	7	Тема 7. Анализ Ссылок	2	2	Проверка лабораторной работы	ПК-3
6	8	Тема 8. Задачи ИП	4	4	Заслушивание доклада, проверка презентации к докладу, ответы на вопросы по теме доклада	ПК-3
		<b>Всего</b>	16	16		

#### 4.3.2. Перечень тем (вопросов), выносимых на самостоятельное изучение студентами в рамках самостоятельной работы (СР)

Тема	Задание	Формируемые компетенции
Тема 2. Булева модель ИП	Решить задачи для самостоятельного выполнения	ПК-3
Тема 3. Лексикон и списки словопозиций	Решить задачи для самостоятельного выполнения	ПК-3
Тема 5. Сжатие индекса	Решить задачи для самостоятельного выполнения	ПК-3

Тема 6. Векторная модель ИП	Решить задачи для самостоятельного выполнения	ПК-3
Тема 7. Анализ Ссылок	Решить задачи для самостоятельного выполнения	ПК-3
Тема 8. Задачи ИП	Подготовить доклад по выбранному варианту	ПК-3

#### 4.4. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ОРГАНИЗАЦИИ САМОСТОЯТЕЛЬНОЙ РАБОТЫ СТУДЕНТОВ

Во время изучения дисциплины студент посещает лекции, практические занятия, выполняет лабораторные задания, готовится к тестам, зачетам и экзаменам. Для каждого вида деятельности необходимо правильно организовать самостоятельную работу.

Лекции. В высшем учебном заведении лекция является важной формой учебного процесса. На лекции студенты получают глубокие и разносторонние знания. Лекция способствует развитию творческих способностей, формирует идейную убежденность, позволяет устанавливать связь учебного материала с производством, новейшими научными достижениями. Лекция требует три вида деятельности: подготовку к лекции, работу на лекции и работу после лекции.

После прослушивания лекции студент должен проработать и осмыслить полученный материал. На каждый пример, приведенный на лекции, желательно, (если это возможно) привести свой. Материал, изложенный в лекции, можно просмотреть в других источниках.

В процессе лекционного занятия студент должен выделять важные моменты, выводы, анализировать основные положения. Недостаточно только «слушать» лекцию. Возможности памяти человека не универсальны. Как бы внимательно студент ни слушал лекцию, большая часть информации вскоре после восприятия будет забыта. Чтобы восстановить лекционный материал, его нужно повторить, а для этого лекцию необходимо конспектировать. Конспект лекций должен быть в отдельной тетради, в которой не должно быть ничего, кроме лекции. Не надо стремиться подробно слово в слово записывать всю лекцию. Конспектируйте только самое важное в рассматриваемой теме: ключевые слова и их значения, примеры использования конструкций, что старается выделить лектор, на чем акцентирует внимание студентов.

Тетрадь для конспекта лекций также требует особого внимания. Ее нужно сделать удобной, практичной и полезной, ведь именно она является основным информативным источником при подготовке к различным отчетным занятиям, зачетам, экзаменам. Конечно, оформление лекционной тетради – это дело вкуса. Но целесообразно отделить поля, где студент мог бы изложить свои мысли, вопросы, появившиеся в ходе лекции. Полезно одну из страниц оставлять свободной. Она потребуется потом, при самостоятельной подготовке. Сюда можно будет занести дополнительную информацию по данной теме, полученную из других источников: рисунки, схемы, примеры кода и т.д.

Лабораторное занятие. Лабораторные занятия по решению задач существенно дополняют лекции. В процессе анализа и решения задач студенты расширяют и углубляют знания, полученные из лекционного курса и учебников, приобретают умение применять общие закономерности к конкретным случаям.

Необходимо, чтобы студенты готовили теоретический материал, т.к. именно невыполнение этого требования приводит к неудаче при решении задач.

Несмотря на различие в видах задач, их решение можно проводить по следующему общему плану (некоторые пункты плана могут выпадать в некоторых конкретных случаях): а) прочесть внимательно условие задачи; б) посмотреть, все ли термины в условиях задачи известны и понятны (если что-то неясно, следует обратиться к учебнику, просмотреть решения предыдущих задач, посоветоваться с

преподавателем); в) произвести анализ задачи, (нужно чётко понимать, в чем будет заключаться решение задачи); г) решить задачу; д) протестировать полученное решение на данных из примеров к задаче, а также на дополнительных данных.

Если задача не решена или «не решается», то необходимо еще раз вернуться к пунктам а) и б). Сколько раз нужно возвращаться к этим пунктам? Практика показывает, что не более десяти раз. Если и после этого задача «не решается», то можно попытаться найти решение этой или похожей задачи в различных источниках.

Домашнее задание. При выполнении домашнего задания необходимо просмотреть текст лекции, разобраться с новыми определениями, посмотреть задания, которые были выполнены на лабораторной работе и применить полученные знания для выполнения домашней работы.

Тест. В первую очередь постарайтесь узнать чего ждать от теста, какие примерно там будут задания. Если вам доступны образцы теста (как, например, при сдаче ЕГЭ), необходимо этим воспользоваться и ежедневно тренироваться.

Не оставляйте все на самый последний момент. Если будете постоянно готовиться к тесту, вы наверняка улучшите свои знания. Для этого составьте план на каждый день, чтобы правильно распределять свое время.

Делайте небольшие перерывы во время учебы. В промежутках можно дать себе небольшую физическую нагрузку. Мозг лучше всего работает, когда умственный труд сменяется физическим. Прогуляйтесь, побегайте, поиграйте в баскетбол, попинайте мяч – помимо стимуляции умственной деятельности, это снимет стресс.

Отдых и контроль над волнением — одни из главных составляющих успеха при подготовке к тесту. Часто ошибки совершаются только из-за стресса, который мешает сконцентрироваться и собраться. Чтобы быть отдохнувшим и расслабленным, соблюдайте составленный режим и старайтесь высыпаться.

Экзамен. На экзамене оцениваются: 1) понимание и степень усвоения теории; 2) методическая подготовка; 3) знание фактического материала; 4) знакомство с обязательной литературой; 5) умение приложить теорию к практике, решать практические задачи и т. д.; 6) логика, структура и стиль ответа, умение защищать выдвигаемые положения. Но значение экзаменов не ограничивается проверкой знаний. Являясь естественным завершением работы студента, они способствуют обобщению и закреплению знаний и умений, приведению их в строгую систему, а также устранению возникших в процессе занятий пробелов.

Студенты готовятся к экзаменам по-разному. Одни из них прорабатывают лишь некоторые вопросы, выбранные наугад, другие стремятся запомнить весь материал подряд, не вникая глубоко в его суть. Работа при этом концентрируется на одном стремлении – сдать экзамен. Недостатки такой системы очевидны. Очевидно также, что подготовка не должна ограничиваться чтением лекционных записей. Первоначальные необработанные конспекты студента содержат факты, определения, выводы, сделанные преподавателем, но в них, как правило, слабо просматривается связующая идея курса, так как студент, записывая каждую лекцию в отдельности, редко способен сразу и достаточно точно уловить общую направляющую мысль. Поэтому конспект требует дополнительной обработки на основе использования учебников и рекомендованной литературы.

Существенные недостатки имеет и такой способ подготовки к экзаменам, как беглый просмотр всего материала. Он эффективен только на некоторых этапах планирования и закрепляющего повторения. Более надежный и целесообразный путь – это тщательная систематизация материала при вдумчивом повторении, установлении внутрисубъектных связей, увязке различных тем и разделов, закреплении путем решения задач.

Перед экзаменом назначается консультация. Цель ее – дать ответы на вопросы, возникшие в ходе самостоятельной подготовки. Хотелось бы обратить особое внимание на важность предэкзаменационных консультаций. Здесь студент имеет полную возможность получить ответ на все неясные ему вопросы. А для этого он должен проработать до консультации весь курс. Кроме того, преподаватель будет отвечать на вопросы других студентов, что будет для вас повторением и закреплением знаний. И еще очень важное обстоятельство: лектор на консультации, как правило, обращает внимание на те разделы, по которым на предыдущих экзаменах ответы были неудовлетворительными, а также фиксирует внимание на наиболее трудных разделах курса. Некоторые студенты не приходят на консультации либо потому, что считают, что у них нет вопросов к лектору, либо полагают, что у них и так мало времени и лучше самому почитать материал по конспекту или в учебнике. Это глубокое заблуждение. Никакая другая работа не сможет принести столь значительного эффекта накануне экзамена, как консультация преподавателя.

Подготовку к экзамену следует начинать с первого дня изучения дисциплины. Как правило, на лекциях подчеркиваются наиболее важные и трудные вопросы или разделы курса, требующие внимательного изучения и обдумывания. Нужно эти вопросы выделить и обязательно постараться разобраться в них, не дожидаясь экзамена, проработать их, готовясь к лабораторным занятиям, попробовать самостоятельно решить несколько типовых задач. И если, несмотря на это, часть материала осталась неувоенной, ни в коем случае нельзя успокаиваться, надеясь на то, что это не попадет на экзамене. Факты говорят об обратном: если те или другие вопросы курса не вошли в экзаменационный билет, преподаватель может их задать (и часто задает) в виде дополнительных вопросов. Точно такое же отношение должно быть выработано к вопросам и задачам, перечисленным в экзаменационной программе, выдаваемой студентам еще до экзамена. Обычно эти же вопросы и аналогичные задачи содержатся в экзаменационных билетах. Не следует оставлять без внимания ни одного раздела курса; если не удалось в чем-то разобраться самому, нужно обратиться к товарищам; если и это не помогло выяснить какой-либо вопрос до конца, нужно обязательно задать этот вопрос преподавателю на предэкзаменационной консультации. Чрезвычайно важно приучить себя к умению самостоятельно мыслить, учиться думать, понимать суть дела. Очень полезно после проработки каждого раздела восстановить в памяти содержание изученного материала, кратко записав это на листе бумаги. Если этого не сделать, то большая часть материала останется не понятой, а лишь формально заученной, и при первом же вопросе экзаменатора студент убедится в том, насколько поверхностно он усвоил материал.

#### **4.5. ПРИМЕРНАЯ ТЕМАТИКА КУРСОВЫХ РАБОТ (ПРОЕКТОВ)**

Не предусмотрено.

## **5. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)**

### **а) перечень литературы**

основная литература:

1. Зинченко А.С. Алгоритмы информационного поиска и их сложность: учебное пособие / А.С. Зинченко. - Иркутск: Изд-во ИГУ, 2019. – 1 электрон. опт. диск (CD-ROM). - ISBN 978-5-9624-1770-7 (ЭБС Библиотех)
2. Зинченко А.С. Алгоритмы информационного поиска и их сложность: практикум / А.С. Зинченко. - Иркутск: Изд-во ИГУ, 2019. – 1 электрон. опт. диск (CD-ROM). - ISBN 978-5-9624-1771-4 (ЭБС Библиотех)

дополнительная литература:

1. Маннинг К.Д. Введение в информационный поиск : научное издание / К. Д. Маннинг, П. Рагхаван, Х. Шютце. - М. : Вильямс, 2011. - 520 с. - ISBN 978-5-8459-1623-5. (15 экз.)

### **б) периодические издания**

### **в) список авторских методических разработок:**

Материалы курса, опубликованные в ИОС «DOMIC».

### **г) базы данных, информационно-справочные и поисковые системы \_\_\_\_\_**

## **6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ**

### **6.1. УЧЕБНО-ЛАБОРАТОРНОЕ ОБОРУДОВАНИЕ:**

Для проведения лекционных занятий необходима аудитория с презентационным оборудованием, для проведения лабораторных занятий необходима аудитория на 25-30 рабочих мест (в зависимости от численности учебной группы), оборудованная доской, презентационной техникой, компьютерами.

### **6.2. ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ:**

1. Комплект разработчика приложений Java Platform (JDK) 11, Standard Edition (распространяется бесплатно);
2. Интегрированная среда разработки NetBeans IDE 12 (распространяется бесплатно, LGPLv2.1, GPLv2 with Classpatch exception).

### **6.3. ТЕХНИЧЕСКИЕ И ЭЛЕКТРОННЫЕ СРЕДСТВА:**

ИОС EDUCA, DOMIC, презентационное оборудование, персональный компьютер с возможностью демонстрации презентаций в формате pdf.

## **7. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ**

При реализации данного курса используются следующие образовательные технологии: технологии традиционного обучения, технологии проблемного обучения, технологии контекстного обучения, интерактивные технологии, технологии дистанционного обучения.

## **8. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ**

Разноуровневые задания к лабораторному практикуму

### **8.1. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ВХОДНОГО КОНТРОЛЯ**

Не предусмотрены.

### **8.2. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ТЕКУЩЕГО КОНТРОЛЯ**

Разноуровневые задания к лабораторному практикуму, доклад.

## Примеры оценочных средств текущего контроля

### 1. Лабораторная работа по теме «Алгоритм PageRank»

Реализуйте алгоритм PageRank для набора данных, представленного в виде списка смежности веб-страниц. Примеры файлов с данными [tiny.txt](#) и [medium.txt](#).

#### Формат ввода

Предполагаем, что существует  $N$  веб-страниц, пронумерованных от 0 до  $N - 1$ , представляем ссылки с упорядоченными парами таких чисел, сначала указывающие страницу, содержащую ссылку, и вторую, указывающую страницу, на которую она ссылается. Формат ввода представляет собой целое число (значение  $n$ ), за которым следует последовательность пар целых чисел (представления всех ссылок).

#### Параметры

Работа алгоритма должна зависеть от параметров:

- **alpha** – вероятность телепортации (например, 0,15);
- **accuracy** – порог стабилизации (например,  $10^{-8}$ ).

Формула для вычисления элементов  $P_{ij}$  матрицы вероятностей переходов  $P$ .


$$P_{ij} = \frac{\alpha}{N} + (1 - \alpha) \cdot p_{ij}$$

где  $N$  – общее количество веб-страниц в графе,  $\alpha$  – коэффициент телепортации,  $p_{ij}$  – вероятность перехода с  $i$ -й страницы на  $j$ -ю (вычисляется как отношение количества ссылок с  $i$ -й страницы на  $j$ -ю к общему количеству исходящих ссылок страницы  $i$ ).

#### Проверка

Для проверки корректности работы алгоритма используйте следующий пример. На наборе данных из файла [tiny.txt](#) должны получиться примерно следующие значения при **alpha** = 0,1:

```
0,27306 0,26586 0,14622 0,24729 0,06757
```

 **Рекомендация.** Для повышения эффективности работы программы используйте алгоритм быстрого возведения в степень.

### 2. Темы докладов

#### Индексирование звука (Shazam)

[http://dmilvdv.narod.ru/Translate/ISASA/An\\_Industrial-Strength\\_Audio\\_Search\\_Algorithm\\_ru.pdf](http://dmilvdv.narod.ru/Translate/ISASA/An_Industrial-Strength_Audio_Search_Algorithm_ru.pdf)

<http://www.redcode.nl/blog/2010/06/creating-shazam-in-java/>

#### Вопросно-ответные системы

[http://ru.wikipedia.org/wiki/Вопросно-ответная\\_система](http://ru.wikipedia.org/wiki/Вопросно-ответная_система)

<http://yury.name/internet/06ia-seminar-note.doc>

<http://rykov-qa2.narod.ru/>

#### Рекомендующие системы. Совместная фильтрация

[http://ru.wikipedia.org/wiki/Коллаборативная\\_фильтрация](http://ru.wikipedia.org/wiki/Коллаборативная_фильтрация)

[http://ru.wikipedia.org/wiki/Рекомендательные\\_системы](http://ru.wikipedia.org/wiki/Рекомендательные_системы)

[http://community.livejournal.com/ru\\_preference/](http://community.livejournal.com/ru_preference/)

<http://yury.name/internet/07ia-seminar-note.pdf>

#### Поиск в полуструктурированных данных

<http://skif.pereslavl.ru/~abram/share/CD-ROM-final/e-book/e-book/1-2/04-Vasenin-Poisk-p-211.pdf>

<http://www.searchtools.com/info/xml.html>

<http://yury.name/internet/08ia-seminar-note.doc>

#### Индексирование текста для поиска с учетом орфографических ошибок

<http://www14.in.tum.de/personen/maass/TUM-I0503.pdf>

<http://yury.name/internet/09ia-seminar-note.doc>

## 8.3. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПРОМЕЖУТОЧНОГО КОНТРОЛЯ

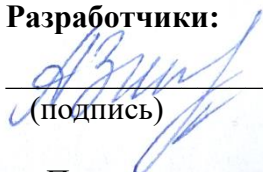
### Список вопросов для промежуточной аттестации:

#### Вопросы для самоконтроля

1. Проблемы развития Интернета. Структурированные и неструктурированные данные. Термин информационный поиск. Запрос. Объект запроса.
2. Парадигмы поиска. Этапы информационного поиска. Виды поиска.
3. Булева модель информационного поиска. Создание инвертированного индекса. Обработка булевых запросов.
4. Булева модель информационного поиска. Сравнение расширенной булевой модели и ранжированного поиска.

5. Лексикон и списки словопозиций. Схематизация документа и декодирование последовательности символов. Определение лексикона терминов.
6. Быстрое пересечение инвертированных списков с помощью указателей пропусков.
7. Словопозиции с координатами и фразовые запросы.
8. Статистические характеристики термов в информационном поиске. Сжатие словаря.
9. Сжатие инвертированного файла.
10. Параметрические и зонные индексы. Частота термина и взвешивание.
11. Модель векторного пространства для ранжирования.
12. Варианты функции tf-idf.
13. Веб как граф. Порталы и авторитетные источники.
14. Алгоритм PageRank.
15. Алгоритм HITS.

**Разработчики:**

  
(подпись)

\_\_\_\_\_  
доцент  
(занимаемая должность)

\_\_\_\_\_  
Зинченко А.С.  
(Ф.И.О.)

Программа составлена в соответствии с требованиями ФГОС ВО по направлению подготовки 09.03.03 «Прикладная информатика» (уровень бакалавриата), утвержденный приказом Министерства образования и науки Российской Федерации от «19» сентября 2017 г. № 922, зарегистрированный в Минюсте России «12» октября 2017 г. № 48531 с изменениями и дополнениями от 26.11.2020 г., 8.02.2021 г.

Программа рассмотрена на заседании кафедры Алгебраических и информационных систем ИМИТ ИГУ «29» апреля 2021 г.

Протокол № 11 Зав. кафедрой  Пантелеев В.И.

*Настоящая программа, не может быть воспроизведена ни в какой форме без предварительного письменного разрешения кафедры-разработчика программы.*