МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования "Пермский государственный национальный исследовательский университет"

Кафедра математического обеспечения вычислительных систем

Авторы-составители: **Ланин Вячеслав Владимирович**

Рабочая программа дисциплины

АЛГОРИТМИЧЕСКИЕ МЕТОДЫ СТРУКТУРНОГО АНАЛИЗА ДОКУМЕНТОВ

Код УМК 83154

Утверждено
Протокол №5
от «09» июня 2020 г.

Пермь, 2020

1. Наименование дисциплины

Алгоритмические методы структурного анализа документов

2. Место дисциплины в структуре образовательной программы

Дисциплина входит в вариативную часть Блока « Б.1 » образовательной программы по направлениям подготовки (специальностям):

Направление: **01.03.02** Прикладная математика и информатика

направленность Интеллектуальный анализ данных и математическое моделирование

3. Планируемые результаты обучения по дисциплине

В результате освоения дисциплины **Алгоритмические методы структурного анализа документов** у обучающегося должны быть сформированы следующие компетенции:

01.03.02 Прикладная математика и информатика (направленность : Интеллектуальный анализ данных и математическое моделирование)

ПК.5 способность осуществлять целенаправленный поиск информации о новейших научных и технологических достижениях в сети Интернет и из других источников

ПК.7 способность к разработке и применению алгоритмических и программных решений в области системного и прикладного программного обеспечения

4. Объем и содержание дисциплины

Направления подготовки	01.03.02 Прикладная математика и информатика (направленность: Интеллектуальный анализ данных и математическое моделирование)
форма обучения	очная
№№ триместров, выделенных для изучения дисциплины	11
Объем дисциплины (з.е.)	3
Объем дисциплины (ак.час.)	108
Контактная работа с преподавателем (ак.час.), в том числе:	42
Проведение лекционных занятий	14
Проведение практических занятий, семинаров	14
Проведение лабораторных работ, занятий по иностранному языку	14
Самостоятельная работа (ак.час.)	66
Формы текущего контроля	Входное тестирование (1) Защищаемое контрольное мероприятие (4)
Формы промежуточной аттестации	Зачет (11 триместр)

5. Аннотированное описание содержания разделов и тем дисциплины

Алгоритмические методы структурного анализа документов

Одним из важнейших ресурсов современной организации является информация, в качестве носителя которой наиболее часто выступают документы. Все виды деятельности организации выражаются посредством тех или иных документов: приказов, служебных записок, договоров, счетов, накладных, личных дел сотрудников и т.д. Очевидно, что чем больше организация, тем сложнее управлять потоком документов.

Настоящее время характеризуется постепенным переходом от бумажных носителей информации к электронным. По оценкам аналитиков уже через 10 лет около 90% документов будут представлены только в электронном виде. Таким образом, место современных бумажных документов вскоре могут занять их электронные эквиваленты. Следует заметить, что полная аналогия неуместна. Например, для электронного документа не существует понятия оригинала. Для работы с электронными документами необходимы свои методы работы,

Общие вопросы обработки электронных документов

Общие вопросы обработки электронных документов

Понятие документа

Традиционный документ. Определение понятия «документ». Унификация и стандартизация документов. Отличия электронного документа от традиционного. Концепция современного электронного документа. Системы управления документами.

Функции и свойства документа

Свойства документа: атрибутивность документа, функциональность документа. Отличительные признаки документа: наличие смыслового семантического содержания, стабильная вещественная форма, предназначенность для использования в социальной коммуникации, завершенность сообщения. .

Проблемы использования электронных документов

Отсутствие семантики в современных электронных документах. Юридические проблемы. Проблемы безопасности. Проблемы достоверности информации.

Форматы электронных документов

Обычный текст. Формат HTML. Формат PDF. Формат XPS (XML Paper Specification). Би-нарные форматы Microsoft Office. Open XML. Open Document Format.

Задачи обработки электронных документов

Задачи обработки электронных документов

Статистический анализ текстов

Задача выделения ключевых слов. Гиперболические законы текста. Закон Ципфа. Закон Ципфа. Вероятность встречи слова в тексте. Канонический закон Ципфа. Диапазон ключевых слов. Понятие относительной частоты. Обратная документная частота. Оценка различительной силы термина.

Индексирование текстовых документов

Блочное индексирование, основанное на сортировке. Однопроходное индексирование в оперативной памяти. Распределенное индексирование. Динамическое индексирование. Статистические характеристики терминов в информационном поиске. Сжатие словаря. Сжатие инвертированного файла.

Модели поиска электронных документов

Булева модель поиска: классическая булева модель, расширенная булева модель, модель нечеткого

поиска. Векторно-пространственная модель поиска. Вероятностная модель поиска. Поиск в пиринговых сетях. Информационно-поисковые языки. Характеристики информационного поиска.

Классификация и каталогизация документов

Классификация и каталогизация документов

Автоматическое реферирование текстовых документов

Задача автоматического реферирования. Виды рефератов. Направления квазиреферирования. Определение веса фрагментов при квазиреферирования. Формирование краткого изложения. Методы оценки.

Технологии Semantic Web

Основные тенденции развития интернет-технологий.
Описание ресурсов на языке RDF. Язык описания онтологий OWL. Стандартны представления метаданных. Технология FOAF.
Интеллектуальные агенты и мультиагентные технологии. алгоритмы обработки данных в Semantic Web.

Технология Wiki

Общие сведения и преимущества. Инструменты создания Вики: MediaWiki, TikiWiki, UseModWiki, FlexWiki. Вики-разметка. Семантическая Вики. Типизированные ссылки. Атрибуты. Основные преимущества Семантической Википедии. Опыт внедрения.

6. Методические указания для обучающихся по освоению дисциплины

Освоение дисциплины требует систематического изучения всех тем в той последовательности, в какой они указаны в рабочей программе.

Основными видами учебной работы являются аудиторные занятия. Их цель - расширить базовые знания обучающихся по осваиваемой дисциплине и систему теоретических ориентиров для последующего более глубокого освоения программного материала в ходе самостоятельной работы. Обучающемуся важно помнить, что контактная работа с преподавателем эффективно помогает ему овладеть программным материалом благодаря расстановке необходимых акцентов и удержанию внимания интонационными модуляциями голоса, а также подключением аудио-визуального механизма восприятия информации.

Самостоятельная работа преследует следующие цели:

– закрепление и совершенствование теоретических знаний, полученных на лекционных занятиях;

- формирование навыков подготовки текстовой составляющей информации учебного и научного назначения для размещения в различных информационных системах;

- совершенствование навыков поиска научных публикаций и образовательных ресурсов, размещенных в сети Интернет;

- самоконтроль освоения программного материала.

Обучающемуся необходимо помнить, что результаты самостоятельной работы контролируются преподавателем во время проведения мероприятий текущего контроля и учитываются при промежуточной аттестации.

Обучающимся с ОВЗ и инвалидов предоставляется возможность выбора форм проведения мероприятий текущего контроля, альтернативных формам, предусмотренным рабочей программой дисциплины. Предусматривается возможность увеличения в пределах 1 академического часа времени, отводимого на выполнение контрольных мероприятий.

Процедура оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья по дисциплине предусматривает предоставление информации в формах, адаптированных к ограничениям их здоровья и восприятия информации.

При проведении текущего контроля применяются оценочные средства, обеспечивающие передачу информации, от обучающегося к преподавателю, с учетом психофизиологических особенностей здоровья обучающихся.

7. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

При самостоятельной работе обучающимся следует использовать:

– конспекты лекций;

– литературу из перечня основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля);

– текст лекций на электронных носителях;

– ресурсы информационно-телекоммуникационной сети "Интернет", необходимые для освоения дисциплины;

– лицензионное и свободно распространяемое программное обеспечение из перечня информационных технологий, используемых при осуществлении образовательного процесса по дисциплине;

– методические указания для обучающихся по освоению дисциплины.

8. Перечень основной и дополнительной учебной литературы

Основная:

1. Федоров, Д. Ю. Программирование на языке высокого уровня Python : учебное пособие для прикладного бакалавриата / Д. Ю. Федоров. — 2-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2019. — 161 с. — (Бакалавр. Прикладной курс). — ISBN 978-5-534-10971-9. — Текст : электронный // ЭБС Юрайт [сайт]. https://www.urait.ru/bcode/437489

2. Загорулько, Ю. А. Искусственный интеллект. Инженерия знаний : учебное пособие для вузов / Ю. А. Загорулько, Г. Б. Загорулько. — Москва : Издательство Юрайт, 2020. — 93 с. — (Высшее образование). — ISBN 978-5-534-07198-6. — Текст : электронный // ЭБС Юрайт [сайт]. https://urait.ru/bcode/455500

Дополнительная:

1. Ланин В. В.,Лядова Л. Н.,Рычков А. Ю. Системы управления электронными документами:учеб.-метод. пособие/В. В. Ланин, Л. Н. Лядова, А. Ю. Рычков.-Пермь:Перм. гос. ун-т,2007, ISBN 5-7944-1023-X.-84.-Библиогр.: с. 82

2. Аналитико-синтетическая переработка информации. Часть 3. Предметизация документов и координатное индексирование документов : учебно-методический комплекс по направлению подготовки 071900 «Библиотечно-информационная деятельность», профилям «Информационно-аналитическая деятельность», «Технолог автоматизированных библиотечно-информационных систем», квалификация (степень) выпускника: бакалавр, форма обучения: очная, заочная / составители О. Я. Сакова. — Кемерово : Кемеровский государственный институт культуры, 2013. — 95 с. — ISBN 2227-8397. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. http://www.iprbookshop.ru/29650

3. Бессмертный, И. А. Интеллектуальные системы : учебник и практикум для академического бакалавриата / И. А. Бессмертный, А. Б. Нугуманова, А. В. Платонов. — Москва : Издательство Юрайт, 2019. — 243 с. — (Бакалавр. Академический курс). — ISBN 978-5-534-01042-8. — Текст : электронный // ЭБС Юрайт [сайт]. https://www.urait.ru/bcode/433716

4. Электронные документы : создание и использование в публичных библиотеках:справочник/О. А. Александрова [и др.] ; науч. ред.: Р. С. Гиляревский, Г. Ф. Гордукалова.-СПб.:Профессия,2007, ISBN 978-5-93913-134-6.-663.-Библиогр. в конце разд.

5. Романенко В. Н.,Никитина Г. В. Сетевой информационный поиск:практ. пособие/Рос. акад. естеств. наук, Сев.-Зап. отд-ние образования и развития науки.-СПб.:Профессия,2003, ISBN 5-93913-044-5.-288.-Библиогр.: с. 284

6. Никитина С. Е. Тезаурус по теоретической и прикладной лингвистике: (Автоматическая обработка текста)/С. Е. Никитина.-Москва:Наука,1978.-376.

7. Леонтьев Б. К. Форматы файлов Microsoft Windows XP:справочник 2005/Б. К. Леонтьев.-М.:Новый издательский дом,2005, ISBN 5-9643-0059-6.-352.

9. Перечень ресурсов сети Интернет, необходимых для освоения дисциплины

http://www.intuit.ru/studies/courses/1064/170/info Математическая теория формальных языков

10. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине

Образовательный процесс по дисциплине **Алгоритмические методы структурного анализа документов** предполагает использование следующего программного обеспечения и информационных справочных систем:
Система программирования Visual Studio Community (бесплатная лицензия для использования в образовательных целях)

При освоении материала и выполнения заданий по дисциплине рекомендуется использование материалов, размещенных в Личных кабинетах обучающихся ЕТИС ПГНИУ (**student.psu.ru**).

При организации дистанционной работы и проведении занятий в режиме онлайн могут использоваться:

система видеоконференцсвязи на основе платформы BigBlueButton (https://bigbluebutton.org/).

система LMS Moodle (http://e-learn.psu.ru/), которая поддерживает возможность использования текстовых материалов и презентаций, аудио- и видеоконтент, а так же тесты, проверяемые задания, задания для совместной работы.

система тестирования Indigo (https://indigotech.ru/).

11. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

Для проведения практических занятий и лабораторных работ необходим компьютерный класс, оборудованный современными рабочими станциями и проектором.

Помещения научной библиотеки ПГНИУ для обеспечения самостоятельной работы обучающихся:

1. Научно-библиографический отдел, корп.1, ауд. 142. Оборудован 3 персональными компьютера с доступом к локальной и глобальной компьютерным сетям.
2. Читальный зал гуманитарной литературы, корп. 2, ауд. 418. Оборудован 7 персональными компьютерами с доступом к локальной и глобальной компьютерным сетям.
3. Читальный зал естественной литературы, корп.6, ауд. 107а. Оборудован 5 персональными компьютерами с доступом к локальной и глобальной компьютерным сетям.
4. Отдел иностранной литературы, корп.2 ауд. 207. Оборудован 1 персональным компьютером с доступом к локальной и глобальной компьютерным сетям.
5. Библиотека юридического факультета, корп.9, ауд. 4. Оборудована 11 персональными компьютерами с доступом к локальной и глобальной компьютерным сетям.
6. Читальный зал географического факультета, корп.8, ауд. 419. Оборудован 6 персональными компьютерами с доступом к локальной и глобальной компьютерным сетям.

Все компьютеры, установленные в помещениях научной библиотеки, оснащены следующим программным обеспечением:

Операционная система ALT Linux;

Офисный пакет Libreoffice.

Справочно-правовая система «КонсультантПлюс»

Фонды оценочных средств для аттестации по дисциплине
Алгоритмические методы структурного анализа документов

Планируемые результаты обучения по дисциплине для формирования компетенции и критерии их оценивания

Компетенция	Планируемые результаты обучения	Критерии оценивания результатов обучения
ПК.5 способность осуществлять целенаправленный поиск информации о новейших научных и технологических достижениях в сети Интернет и из других источников	Студент знает основные методы обработки и семантического анализа текстов. Студент умеет применять семантические технологии при получении, хранении и переработки информации. Студент владеет технологиями интеллектуального поиска в гетерогенных источниках информации.	**Неудовлетворител** Студент не знает основные методы обработки и семантического анализа текстов, не умеет применять семантические технологии при получении, хранении и переработки информации и не владеет технологиями интеллектуального поиска в гетерогенных источниках информации. **Удовлетворительн** Студент знает основные методы обработки и семантического анализа текстов, но не умеет применять семантические технологии при получении, хранении и переработки информации и не владеет технологиями интеллектуального поиска в гетерогенных источниках информации. **Хорошо** Студент знает основные методы обработки и семантического анализа текстов, умеет применять семантические технологии при получении, хранении и переработки информации, но не владеет технологиями интеллектуального поиска в гетерогенных источниках информации. **Отлично** Студент знает основные методы обработки и семантического анализа текстов, умеет применять семантические технологии при получении, хранении и переработки информации и владеет технологиями интеллектуального поиска в гетерогенных источниках информации.
ПК.7 способность к разработке и применению	Студент знает основные подходы к реализации обработки и семантического анализа текстов.	**Неудовлетворител** Студент не знает основные подходы к реализации обработки и семантического анализа текстов, не умеет применять

Компетенция	Планируемые результаты обучения	Критерии оценивания результатов обучения
алгоритмических и программных решений в области системного и прикладного программного обеспечения	Студент умеет применять лингвистические ресурсы и библиотеки при реализации программных систем. Студент владеет инструментальными технологиями реализации обработки и семантического анализа текстов.	**Неудовлетворител** лингвистические ресурсы и библиотеки при реализации программных систем, не владеет инструментальными технологиями реализации обработки и семантического анализа текстов. **Удовлетворительн** Студент знает основные подходы к реализации обработки и семантического анализа текстов, но не умеет применять лингвистические ресурсы и библиотеки при реализации программных систем и не владеет инструментальными технологиями реализации обработки и семантического анализа текстов. **Хорошо** Студент знает основные подходы к реализации обработки и семантического анализа текстов, умеет применять лингвистические ресурсы и библиотеки при реализации программных систем, но не владеет инструментальными технологиями реализации обработки и семантического анализа текстов. **Отлично** Студент знает основные подходы к реализации обработки и семантического анализа текстов. Студент умеет применять лингвистические ресурсы и библиотеки при реализации программных систем. Студент владеет инструментальными технологиями реализации обработки и семантического анализа текстов.

Оценочные средства текущего контроля и промежуточной аттестации

Схема доставки : Базовая

Вид мероприятия промежуточной аттестации : Зачет
Способ проведения мероприятия промежуточной аттестации : Оценка по дисциплине в рамках промежуточной аттестации определяется на основе баллов, набранных обучающимся на контрольных мероприятиях, проводимых в течение учебного периода.
Максимальное количество баллов : 100

Конвертация баллов в отметки

«отлично» - от 81 до 100
«хорошо» - от 61 до 80
«удовлетворительно» - от 57 до 60
«неудовлетворительно» / «незачтено» менее 57 балла

Компетенция	Мероприятие текущего контроля	Контролируемые элементы результатов обучения
Входной контроль	Понятие документа **Входное тестирование**	Письменная работа на знание формальных языков, моделей представления знаний, языков разметки документов, умения составлять программы решения задач обработки текстов
ПК.7 способность к разработке и применению алгоритмических и программных решений в области системного и прикладного программного обеспечения	Форматы электронных документов **Защищаемое контрольное мероприятие**	Выполнение лабораторной работы
ПК.7 способность к разработке и применению алгоритмических и программных решений в области системного и прикладного программного обеспечения	Индексирование текстовых документов **Защищаемое контрольное мероприятие**	реализация программного модуля
ПК.7 способность к разработке и применению алгоритмических и программных решений в области системного и прикладного программного обеспечения	Технологии Semantic Web **Защищаемое контрольное мероприятие**	Разработка онтологии для индексации электронного документа

Компетенция	Мероприятие текущего контроля	Контролируемые элементы результатов обучения
ПК.5 способность осуществлять целенаправленный поиск информации о новейших научных и технологических достижениях в сети Интернет и из других источников **ПК.7** способность к разработке и применению алгоритмических и программных решений в области системного и прикладного программного обеспечения	Технология Wiki **Защищаемое контрольное мероприятие**	Разработка системы Wiki-документов

Спецификация мероприятий текущего контроля

Понятие документа

Продолжительность проведения мероприятия промежуточной аттестации: **1 часа**
Условия проведения мероприятия: **в часы аудиторной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **0**
Проходной балл: **0**

Показатели оценивания	Баллы
Написана программа составления частотного словаря для текстового файла	3
Продемонстрированы знания языков разметки XML, HTML, xHTML	2
Продемонстрировано знание стандарта кодирования Unicode	1
Перечислены задачи, входящие в область информационного поиска	1
Приведена классификация формальных языков	1
Даны определения синтаксиса, семантики и прагматики языка	1
Перечислены известные модели представления знаний	1

Форматы электронных документов

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**
Условия проведения мероприятия: **в часы аудиторной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **25**
Проходной балл: **15**

Показатели оценивания	Баллы
Реализована генерация формата DOCX.	15
Реализована генерация формата XSLX.	10

Индексирование текстовых документов

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**
Условия проведения мероприятия: **в часы аудиторной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **25**
Проходной балл: **12**

Показатели оценивания	Баллы
Реализовано выделение слов текста	8
Реализован подсчет частоты	6
Реализована обработка файлов в формате PDF	6
Реализовано выделение основы слова	5

Технологии Semantic Web

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**
Условия проведения мероприятия: **в часы аудиторной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **25**
Проходной балл: **15**

Показатели оценивания	Баллы
Онтология содержит 10 классов.	15
Онтология содержит 15 классов.	10
Онтология содержит 5 типов отношений.	5

Технология Wiki

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**
Условия проведения мероприятия: **в часы аудиторной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **25**
Проходной балл: **15**

Показатели оценивания	Баллы
Выделены семантические атрибуты и ссылки.	15
Разработано 5 статей.	10