

МИНОБРНАУКИ РОССИИ

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

Кафедра математического обеспечения вычислительных систем

Авторы-составители: **Ланин Вячеслав Владимирович**
Городилов Алексей Юрьевич

Рабочая программа дисциплины

СЕМАНТИЧЕСКИЕ ТЕХНОЛОГИИ ОБРАБОТКИ ТЕКСТОВЫХ ДОКУМЕНТОВ

Код УМК 91494

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

Пермь, 2020

1. Наименование дисциплины

Семантические технологии обработки текстовых документов

2. Место дисциплины в структуре образовательной программы

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

Направление подготовки: **01.03.02** Прикладная математика и информатика
направленность Искусственный интеллект и большие данные

3. Планируемые результаты обучения по дисциплине

В результате освоения дисциплины **Семантические технологии обработки текстовых документов** у обучающегося должны быть сформированы следующие компетенции:

01.03.02 Прикладная математика и информатика (направленность : Искусственный интеллект и большие данные)

ОПК.4 Способен применять и модифицировать математические модели для решения задач в области профессиональной деятельности

Индикаторы

ОПК.4.3 Демонстрирует практический опыт по использованию или модификации готовых математических моделей и моделей данных для решения задач в области профессиональной деятельности

4. Объем и содержание дисциплины

Направления подготовки	01.03.02 Прикладная математика и информатика (направленность: Искусственный интеллект и большие данные)
форма обучения	очная
№№ семестров, выделенных для изучения дисциплины	8
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 (8 семестр)

5. Аннотированное описание содержания разделов и тем дисциплины

Общие вопросы обработки электронных 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.

Понятие документа

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

Регулярные выражения

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

Обработка данных в формате XML и JSON

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). Бинарные форматы Microsoft Office. Open XML. Open Document Format.

Скрапинг веб-сайтов

Работа с Web API

Извлечение данных из социальных сетей

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.

Информационно-поисковые тезаурусы и онтологии

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.

Основные понятия обработки ЕЯ

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 аудиторские 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. Ланин В. В., Лядова Л. Н., Рычков А. Ю. Системы управления электронными документами: учебно-методическое пособие предназначено для студентов, изучающих курс "Системы управления электронными документами"/В. В. Ланин, Л. Н. Лядова, А. Ю. Рычков.-Пермь, 2007, ISBN 5-7944-1023-X.-84.-Библиогр.: с. 82-83
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. Лукашевич Н. В. Тезаурусы в задачах информационного поиска/Н. В. Лукашевич.-Москва:Изд-во Московского ун-та,2011, ISBN 978-5-211-05926-9.-5083.-Библиогр.: с. 483-508
2. Марманис Х.,Бабенко Д. Алгоритмы интеллектуального Интернета:[передовые методики сбора, анализа и обработки данных]/Х. Марманис, Д. Бабенко ; [пер. с англ. М. Низовец].-Санкт-Петербург:Символ-Плюс,2011, ISBN 978-5-93286-186-8.-480.

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 **Семантические технологии обработки текстовых документов** предполагает использование следующего программного обеспечения и информационных справочных систем:

Система программирования Microsoft Visual Studio

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 лекционных занятий требуется аудитория, оснащенная презентационной техникой (проектор, экран, компьютер/ноутбук) с соответствующим программным обеспечением, меловой (и) или маркерной доской.

Для проведения практических занятий - аудитория, оснащенная презентационной техникой (проектор, экран, компьютер/ноутбук) с соответствующим программным обеспечением, меловой (и) или маркерной доской.

Для проведения лабораторных занятий - меловая и (или) маркерная доска, компьютерный класс (аппаратное и программное обеспечение определено в паспортах компьютерных классов).

Для групповых (индивидуальных) консультаций - аудитория, оснащенная меловой (и) или маркерной доской.

Для проведения текущего контроля - аудитория, оснащенная меловой (и) или маркерной доской.

Самостоятельная работа студентов: аудитория, оснащенная компьютерной техникой с возможностью подключения к сети «Интернет», с обеспеченным доступом в электронную информационно-образовательную среду университета, помещения Научной библиотеки ПГНИУ.

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
Семантические технологии обработки текстовых документов**

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

ОПК.4

Способен применять и модифицировать математические модели для решения задач в области профессиональной деятельности

Компетенция (индикатор)	Планируемые результаты обучения	Критерии оценивания результатов обучения
<p>ОПК.4.3 Демонстрирует практический опыт по использованию или модификации готовых математических моделей и моделей данных для решения задач в области профессиональной деятельности</p>	<p>Демонстрирует практический опыт по использованию или модификации готовых математических моделей и моделей данных для решения задач в области обработки текстовых документов</p>	<p align="center">Неудовлетворител</p> <p>Студент не знает и не умеет использовать методы обработки электронных документов. Допускает значительные ошибки. Имеет практический опыт по использованию или модификации готовых математических моделей и моделей данных для решения задач в области обработки текстовых документов, но допускает серьезные ошибки. Студент не знает основные методы обработки и семантического анализа текстов. Студент не умеет применять семантические технологии при получении, хранении и переработки информации. Студент не владеет технологиями интеллектуального поиска в гетерогенных источниках информации</p> <p align="center">Удовлетворительн</p> <p>Студент знает и умеет использовать методы обработки электронных документов. Допускает значительные ошибки. Имеет практический опыт по использованию или модификации готовых математических моделей и моделей данных для решения задач в области обработки текстовых документов.</p> <p align="center">Хорошо</p> <p>Студент знает и умеет использовать методы обработки электронных документов. Допускает незначительные ошибки. Имеет практический опыт по использованию или модификации готовых математических моделей и моделей данных для решения задач в области обработки текстовых документов. Знает не в полном объеме современные тенденции развития интернет-</p>

Компетенция (индикатор)	Планируемые результаты обучения	Критерии оценивания результатов обучения
		<p style="text-align: center;">Хорошо</p> <p>технологий или допускает некоторые ошибки, умеет работать по запросу в рамках своей профессиональной деятельности с поддержкой преподавателя или справочных систем</p> <p style="text-align: center;">Отлично</p> <p>Студент знает и умеет использовать методы обработки электронных документов. Имеет практический опыт по использованию или модификации готовых математических моделей и моделей данных для решения задач в области обработки текстовых документов. Знает современные тенденции развития интернет-технологий, умеет самостоятельно работать по запросу в рамках своей профессиональной деятельности. 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</p>

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 48 до 60

«неудовлетворительно» / «незачтено» менее 48 балла

Компетенция (индикатор)	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
ОПК.4.3 Демонстрирует практический опыт по использованию или модификации готовых математических моделей и моделей данных для решения задач в области профессиональной деятельности	Форматы электронных документов Защищаемое контрольное мероприятие	Выполнение лабораторной работы

Компетенция (индикатор)	Мероприятие текущего контроля	Контролируемые элементы результатов обучения
ОПК.4.3 Демонстрирует практический опыт по использованию или модификации готовых математических моделей и моделей данных для решения задач в области профессиональной деятельности		Реализация программного модуля

Компетенция (индикатор)	Мероприятие текущего контроля	Контролируемые элементы результатов обучения
	Информационно-поисковые тезаурусы и онтологии Защищаемое контрольное мероприятие	
ОПК.4.3 Демонстрирует практический опыт по использованию или модификации готовых математических моделей и моделей данных для решения задач в области профессиональной деятельности	Технологии Semantic Web Защищаемое контрольное мероприятие	Разработка онтологии для индексации электронного документа
ОПК.4.3 Демонстрирует практический опыт по использованию или модификации готовых математических моделей и моделей данных для решения задач в области профессиональной деятельности	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
Приведена классификация формальных языков	1
Продемонстрировано знание стандарта кодирования Unicode	1
Перечислены задачи, входящие в область информационного поиска	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 самостоятельной работы**

Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **25**

Проходной балл: **12**

Показатели оценивания	Баллы
Реализована работа с одним из форматов Open Document	8
Реализована генерация формата DOCX.	6
Реализована генерация формата XSLX.	6
В сгенерированных документах реализовано расширенное форматирование	5

Информационно-поисковые тезаурусы и онтологии

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 основы слова.	10
Реализован подсчет частоты.	7
Реализовано выделение слов текста.	5
Реализована обработка файлов в формате PDF.	3

Технологии Semantic Web

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**

Условия проведения мероприятия: **в часы самостоятельной работы**

Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **25**

Проходной балл: **12**

Показатели оценивания	Баллы
Реализовано 5 запросов на языке SPARQL.	7
Онтология содержит 5 типов отношений.	6
Онтология содержит 10 классов.	6
Онтология содержит 15 экземпляров.	6

Технология Wiki

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**

Условия проведения мероприятия: **в часы самостоятельной работы**

Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **25**

Проходной балл: **12**

Показатели оценивания	Баллы
Выделены 7 семантических атрибутов.	7
Разработано 5 статей.	6
Реализовано 5 запросов на языке SPARQL.	6
Выделены 5 типизированных ссылок.	6