МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
"Российский государственный гуманитарный университет"
(ФГБОУ ВО "РГГУ")

ИНСТИТУТ ИНФОРМАЦИОННЫХ НАУК И ТЕХНОЛОГИЙ БЕЗОПАСНОСТИ
ФАКУЛЬТЕТ ИНФОРМАЦИОННЫХ СИСТЕМ И БЕЗОПАСНОСТИ
КАФЕДРА ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И СИСТЕМ

МЕТОДЫ ОБРАБОТКИ ТЕКСТОВ В ЗАДАЧАХ ИНФОРМАТИЗАЦИИ ГУМАНИТАРНОЙ СФЕРЫ

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

По направлению подготовки 09.03.03 Прикладная информатика
профиль: Прикладная информатика в гуманитарной сфере
Уровень квалификации выпускника бакалавр

Форма обучения очная

РПД адаптирована для лиц
с ограниченными возможностями
здоровья и инвалидов

Москва 2021

МЕТОДЫ ОБРАБОТКИ ТЕКСТОВ В ЗАДАЧАХ ИНФОРМАТИЗАЦИИ ГУМАНИТАРНОЙ СФЕРЫ
Рабочая программа дисциплины
Составитель:
старший преподаватель Охапкина Е.П.

Ответственный редактор
кандидат технических наук, доцент,
зав. кафедрой информационных технологий и систем
А.А. Роганов

УТВЕРЖДЕНО
Протокол заседания кафедры ИТС
№ 12 от 28.06.2021 г.

ОГЛАВЛЕНИЕ

1. Пояснительная записка

1.1. Цель и задачи дисциплины

***Цель* дисциплины:** формирование у студентов научного подхода к освоению, созданию и использованию в гуманитарной сфере интеллектуальных информационных систем, основанных на текстовых базах знаний и естественно-языковых средствах коммуникации.

Задачи: освоить общие принципы построения систем обработки текстов; раскрыть структуру лингвистических процессоров и модульный принцип их построения; освоить методы построения модулей лингвистических процессоров (графематического, морфологического, синтаксического); ознакомиться с принципами и методами построения модуля семантического анализа; освоить основы компьютерной лексикографии; дать представление о возможностях автоматического создания текстов.

1.2. Формируемые компетенции, соотнесённые с планируемыми результатами обучения по дисциплине:

Коды компетенции	**Содержание компетенций**	**Перечень планируемых результатов обучения по дисциплине**
ПК-3 Способен проектировать информационные системы по видам обеспечения	ПК-3.1 Знает модели жизненного цикла информационных систем, основные технологии, стадии и этапы их проектирования	***знать:*** *общие* принципы построения систем автоматической обработки текстов (письменного и устного). ***уметь:*** работать с автоматическими словарями. ***обладать навыками*** работы с инструментами автоматической обработки естественного языка.
	ПК-3.2 Умеет применять технологии проектирования информационных систем по видам обеспечения	***Знать:*** структуру систем синтеза и анализа. ***Уметь:*** охарактеризовать принципы морфологического, синтаксического анализа. ***обладать навыками*** анализа предметных областей для формирования требований к системам автоматической обработки естественного языка.
	ПК-3.3 Владеет навыками проектирования информационных систем или их частей по видам обеспечения	***Знать:*** этапы и уровни автоматической обработки текста; построение графематического, морфологического и синтаксического анализа; практические возможности современных систем автоматической обработки естественного языка.

		Уметь: осуществлять реферирование текста с использованием компьютерных средств. ***обладать навыками*** проектирования систем автоматической обработки естественного языка.

1.3. Место дисциплины в структуре образовательной программы.

Дисциплина «Методы обработки текстов в задачах информатизации гуманитарной сферы» является элективной дисциплиной блока Б1 учебного плана по направлению подготовки 09.03.03 Прикладная информатика Профиль: Прикладная информатика в гуманитарной сфере. Дисциплина реализуется на факультете Информационных систем и безопасности кафедрой информационных технологий и систем. Для освоения дисциплины необходимы знания, умения и владения, сформированные в ходе изучения следующих дисциплин: Информатика, Методы и технологии искусственного интеллекта, Интеллектуальные информационные системы.

В результате освоения дисциплины формируются знания, умения и владения, необходимые для изучения следующих дисциплин: Технологии Big Data, Центры обработки данных.

2. Структура дисциплины

Общая трудоёмкость дисциплины составляет 3 з.е., 114 ч., в том числе контактная работа обучающихся с преподавателем 42 ч., в том числе лекции 14 ч., практические работы 28 ч., самостоятельная работа обучающихся 72 ч.

№ п/п	Раздел дисциплины/темы	Семестр	Виды учебной работы (в часах)						Формы текущего контроля успеваемости, форма промежуточной аттестации *(по семестрам)*
			Контактная					Самостоятельная работа	
			Лекции	Семинар	Практические занятия	Лабораторные занятия	Промежуточная аттестация		
1.	Введение; назначение и области применения естественно-языковых (ЕЯ) систем; предмет и содержание курса.	4	1		4			12	Опрос
2.	Основы семиотики: семиотика как наука	4	1		4			14	Защита отчета по

	о знаках; языковая семиотика; язык как система								практической работе № 1
3.	Науки о знаниях: лингвистическая модель коммуникации - звуки, фонемы, морфемы, слова, предложения; способы хранения знаний; понятие онтологии как статического знания	4	4		4			14	Защита отчетов по практической работе № 2-4
4.	Модели естественного языка: проблемы понимания текстов на ЕЯ; Проблема моделирования языка. Принципы и методы автоматической обработки текстов на ЕЯ	4	4		8			16	Защита отчетов по практическим работам № 5,6
5.	Прикладные естественно-языковые системы: Формы общения – интерфейсы, Системы машинного перевода, Системы общения с текстовыми базами данных	4	4		8			16	Защита отчета по практической работе № 7,8
	Зачет								зачет по билетам
	итого:		14		28			72	

3. Содержание дисциплины

№	Наименование раздела дисциплины	Содержание
1.	Введение; назначение и области применения естественно-языковых (ЕЯ) систем; предмет и содержание курса.	**Введение** Назначение и области применения естественно-языковых (ЕЯ) систем. Предмет и содержание курса. Характеристика, основных проблем создания и совершенствования ЕЯ систем. Автоматизированная обработка текстов на ЕЯ и искусственный интеллект. Основные классы ЕЯ систем: интеллектуальные вопросно-ответные системы, системы общения с базами данных, с экспертными системами, системы обработки связных текстов, системы машинного перевода. **Раздел 1. Основы семиотики** Тема 1.1. Семиотика как наука о знаках Понятие и определение знака. Предметное и смысловое значение знака. Денотат концепт и референт. Треугольник Фреге. Экстенсионал и интенсионал знака. Виды знаков. Семиотические системы. Синтактика, семантика и прагматика. Знак как унитарный носитель информации о факте культуры. Государственные символы. Франчайзинг как коммерческая продажа знака. Текст как сложное знаковое единство. Устные, письменные, печатные тексты. Тексты СМИ. Гипертекст. Интернет как новая среда обитания текстов. Тема 1.2. Языковая семиотика. Символы, индексы, иконические знаки. Морфемы естественного языка как простейшие знаки. Слово как сложный знак. Мотивированность языкового знака Тема 1.3. Язык как система. Понятие знаковой системы. Три аспекта изучения знаковых систем: семантика, синтактика, прагматика. Типология знаковых систем. Текст, смысл, информация. Текст как модель знаковой системы. Свойства и отношения. Операции над отношениями. Свойства отношений. Основные типы отношений в текстах: эквивалентность, толерантность, порядок, древесный порядок. Нективная и юнктивная связи. Правильные, семантичные и осмысленные тексты. Язык и речь как код и сообщение. Парадигматические и синтагматические отношения в языке и речи. Многоуровневая структура языка. Слово в тексте.
2.	Основы семиотики: семиотика как наука о знаках; языковая семиотика; язык как система	**Раздел 2. Науки о знаниях. Общее понятие об онтологии. Онтологизация Интернета.** Тема 2.1. Лингвистическая модель коммуникации - звуки, фонемы, морфемы, слова, предложения. Морфемы - грамматические и смысловые значения. Аналитические и флективные языки. Предложение как минимальный носитель знания.

		Тема 2.2. Способы хранения знаний. Текстовая форма. Способы хранения знаний в машине. Элементарные формы хранения знаний - предикаты, фреймы, нейросети. Представление знаний в машинной форме - предикаты (Аристотель), фреймы (Марвин Минский), концептуальные графы (Джон Сова). Знание в виде сети IF-THEN-ELSE (расширение концепции аристотелева силлогизма). Тема 2.3. Понятие онтологии как статического знания. Онтологии первого и второго уровней. Язык XML и кодирование онтологий. Онтология - набор понятий, связи между ними как набор ролей и отношений. Языки кодирования онтологий. Тема 2.4. Извлечение знаний - knowledge acquisition. Извлечение знаний в кибернетическую систему (AI) - процессы R-AI, M-AI, S-AI. Особенности этих процессов. Язык представления знаний KNOW. Язык (формат) обмена знаниями (KIF). Кодирование простейших единиц знания - отдельных слов, понятий. Как они объединяются в элементарные структуры. Концептуальные графы Д. Совы. Тема 2.5. Управление знаниями в бизнесе (KM2B). KM2B - это технология, включающая в себя комплекс формализованных методов по: поиску и извлечению знаний, структурированию и систематизации знаний, анализу знаний, обновлению (актуализации) знаний, распространению знаний, генерации новых знаний. Базовые характеристики знания в KM2B: - содержательный компонент (идея и контекст применения) - актуальность (знания должны быть "живыми" - сохранять полезность для субъекта)- отчуждаемость, - повторяемость результатов использования знаний при использовании их другими людьми. Тема 2.6. Виды знаний в организации - невыявленные, выявленные, но не отчужденные (записки), выявленные и отчужденные. Способы отчуждения, выявления знаний в производственном коллективе и вовлечения их в коллективное пользование. Как коллектив организации умеет работать со знаниями, генерировать новые.
3.	Науки о знаниях: лингвистическая модель коммуникации - звуки, фонемы, морфемы, слова, предложения; способы хранения знаний; понятие	**Раздел 3. Модели естественного языка** Тема 3.1. Проблемы понимания текстов на ЕЯ Процессы и структуры понимания. Модели памяти. Проблемы понимание текста. Препозиционная репрезентация текста. Модель понимания Кинча. Тема 3.2. Проблема моделирования языка Модель «Непосредственно-составляющих»:

	онтологии как статического знания	особенности, достоинства, недостатки. Модель «Трансформационная порождающая грамматика»: особенности, достоинства , недостатки. Модель «Контрастивная порождающая грамматика», Модель «Падежная грамматика». Модель «Смысл – Текст». Лексическая функция. Тема 3.3. Принципы и методы автоматической обработки текстов на ЕЯ Морфологический, синтаксический, семантический и прагматический анализ и синтез предложений на естественном языке. Базы знаний о языке общения.
4.	Модели естественного языка: проблемы понимания текстов на ЕЯ; Проблема моделирования языка. Принципы и методы автоматической обработки текстов на ЕЯ	**Раздел 4. Технология создания прикладных систем автоматизированной обработки текстов на естественном языке** Тема 4.1.Обобщенная схема ЕЯ системы Основные компоненты интеллектуально-диалоговой системы. Анализ высказываний. Интерпретация высказываний. Генерация смысла высказываний. Синтез высказываний. Диалоговая компонента. Понятие сценария диалога. Методы представления знаний о диалоге. Тема 4.2. Компоненты понимания и генерации высказываний Анализ слов. Методы морфологического анализа. Частотные методы. Полиграммные методы. Абсолютные методы. Относительные методы. Типы анализаторов: традиционные, концептуальные, использующие сопоставление по образцам, смешанные. Концептуальный синтез. Лингвистический синтез. Алгоритмы морфологического, синтаксического и семантического анализа текстов.
5.	Прикладные естественно-языковые системы: Формы общения – интерфейсы, Системы машинного перевода, Системы общения с текстовыми базами данных	**Раздел 5. Прикладные естественно-языковые системы** Тема 5.1. Формы общения – интерфейсы Стандартные интерфейсы. Табличная форма общения. Текстовая форма общения. Лингвистический процессор диалоговой системы. Тема 5.2. Системы машинного перевода Назначение систем машинного перевода (СМП). Классификация СМП. Промышленные, развивающиеся и экспериментальные СМП. Методы представления знаний о естественном языке в компьютерных системах машинного перевода. Тема 5.3. Системы общения с текстовыми базами данных Приобретение знаний из текстов. Модели приобретения знаний. Уровни знаний. Гипертекстовые системы. Методы создания и навигации по гипертексту.

4. Образовательные технологии

№ п/п	Наименование раздела	Виды учебных занятий	Образовательные технологии
1	2	3	4
1.	Введение; назначение и области применения естественно-языковых (ЕЯ) систем; предмет и содержание курса.	Лекции Практическая работа № 1. Самостоятельная работа	Вводная лекция с использованием видеоматериалов Прием отчетов по практической работе № 1 Консультирование по пройденному учебному материалу
2.	Основы семиотики: семиотика как наука о знаках; языковая семиотика; язык как система	Лекции Практическая работа № 2. Самостоятельная работа	Лекция с использованием видеоматериалов Прием отчета по практической работе № 2 Консультирование по пройденному учебному материалу
3.	Науки о знаниях: лингвистическая модель коммуникации - звуки, фонемы, морфемы, слова, предложения; способы хранения знаний; понятие онтологии как статического знания	Лекции Практическая работа № 3. Самостоятельная работа	Вводная лекция с использованием видеоматериалов Прием отчетов по практической работе № 3 Консультирование по пройденному учебному материалу
4.	Модели естественного языка: проблемы понимания текстов на ЕЯ; Проблема моделирования языка. Принципы и методы автоматической обработки текстов на ЕЯ	Лекции Практическая работа № 4. Самостоятельная работа	Лекции с использованием видеоматериалов Прием отчета по практической работе № 4 Консультирование по пройденному учебному материалу
5.	Прикладные естественно-языковые системы: Формы общения – интерфейсы, Системы машинного перевода, Системы общения с текстовыми базами данных	Лекции Практическая работа № 5 Самостоятельная работа	Лекция с использованием видеоматериалов. Прием отчета по практической работе № 5 Консультирование по пройденному учебному материалу

5. Оценка планируемых результатов обучения

5.1. Система оценивания

Форма контроля	**Макс. количество баллов**	
	За одну работу	**Всего**
Текущий контроль: Практическая работа № 1, защита отчета Практическая работа № 2, защита отчета Практическая работа № 3,4, защита отчета Практическая работа № 5,6, защита отчета Практическая работа № 7,8, защита отчета	 12 баллов 12 баллов 12 баллов 12 баллов 12 баллов	 12 баллов 12 баллов 12 баллов 12 баллов 12 баллов
Промежуточная аттестация *зачет*		40 баллов
Итого за семестр		100 баллов

Полученный совокупный результат конвертируется в традиционную шкалу оценок и в шкалу оценок Европейской системы переноса и накопления кредитов (European Credit Transfer System; далее – ECTS) в соответствии с таблицей:

100-балльная шкала	Традиционная шкала		Шкала ECTS
95 – 100	Отлично	зачтено	A
83 – 94			B
68 – 82	Хорошо		C
56 – 67	Удовлетворительно		D
50 – 55			E
20 – 49	Неудовлетворительно	не зачтено	FX
0 – 19			F

5.2.Критерии выставления оценки по дисциплине

Баллы/ Шкала ECTS	**Оценка по дисциплине**	**Критерии оценки результатов обучения по дисциплине**
100-83/ A,B	«зачтено (отлично)»	Выставляется обучающемуся, если он глубоко и прочно усвоил теоретический и практический материал, может продемонстрировать это на занятиях и в ходе промежуточной аттестации. Обучающийся исчерпывающе и логически стройно излагает учебный материал, умеет увязывать теорию с практикой, справляется с решением задач профессиональной направленности высокого уровня сложности, правильно обосновывает принятые решения. Свободно ориентируется в учебной и профессиональной литературе. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «высокий».

82-68/ C	«зачтено (хорошо)»	Выставляется обучающемуся, если он знает теоретический и практический материал, грамотно и по существу излагает его на занятиях и в ходе промежуточной аттестации, не допуская существенных неточностей. Обучающийся правильно применяет теоретические положения при решении практических задач профессиональной направленности разного уровня сложности, владеет необходимыми для этого навыками и приёмами. Достаточно хорошо ориентируется в учебной и профессиональной литературе. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «хороший».
67-50/ D,E	«зачтено (удовлетвори-тельно)»	Выставляется обучающемуся, если он знает на базовом уровне теоретический и практический материал, допускает отдельные ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает определённые затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, владеет необходимыми для этого базовыми навыками и приёмами. Демонстрирует достаточный уровень знания учебной литературы по дисциплине. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «достаточный».
49-0/ F,FX	«не зачтено (неудовлетворите льно)»	Выставляется обучающемуся, если он не знает на базовом уровне теоретический и практический материал, допускает грубые ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает серьёзные затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, не владеет необходимыми для этого навыками и приёмами. Демонстрирует фрагментарные знания учебной литературы по дисциплине. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции на уровне «достаточный», закреплённые за дисциплиной, не сформированы.

5.3. Оценочные средства (материалы) для текущего контроля успеваемости, промежуточной аттестации обучающихся по дисциплине

Контрольные вопросы зачета (ПК-3)

1. Понятие «текст» в информатике. Смысл текста.
2. Знак. Знаковая ситуация. Знаковая система.
3. Три составные части семиотики.
4. Понятия: концепт. Денотат. Треугольник Фреге.
5. Основы формализации текста. Текст как «модель теории».
6. Устройство естественного языка как знаковой системы.
7. Единицы естественного языка в трех компонентах естественного языка.
8. Модель ЕЯ «Непосредственные составляющие».
9. Модель ЕЯ «Порождающие грамматики».
10. Модель ЕЯ «Трансформационные грамматики».
11. Модель ЕЯ «Смысл – текст».
12. Проблема понимания ЕЯ.
13. Модель понимания Кинча.
14. Роль знаний о предметной области в понимании текстов на ЕЯ.
15. Продукционная модель представления знаний.
16. Представление знаний с помощью фреймов.
17. Семантическая сеть – модель представления знаний.
18. Представление знаний средствами логики предикатов.
19. Нечеткие знания и способы их представления в ЭВМ.
20. Сравнительная характеристика моделей представления знаний.
21. Гипертекстовая технология представления знаний.
22. Классификация интеллектуальных диалоговых систем.
23. Обобщенная схема ИДС.
24. Компонента «Ведения диалога»: основные задачи, подходы к их реализации.
25. Компонента «Понимание высказываний»: основные задачи, подходы к их реализации.
26. Компонента «Генерация высказываний»: основные задачи, подходы к их реализации.
27. Лингвистическая модель языка общения.

Тематика рефератов (докладов) (ПК-3)

1. 1. Теоретические основы ЛО. Общая модель коммуникативного взаимодействия. Лингвистические и психологические аспекты коммуникации
2. Теоретические основы ЛО. Основные понятия семиотики и логики,
3. Теоретические основы ЛО Основные понятия теории речевых актов. Пресуппозиция.
4. Классификация информационных систем. Электронные библиотеки как специфический вид АИС
5. Основы теории информационного поиска
6. Семантические языки разметки текста.
7. Различные подходы к определению ЛО. История разработки ЛО в России
8. Классификация средств ЛО.
9. Общие понятия и основные системы метаданных.
10. Языки библиографических данных. Дублинское ядро метаданных.
11. Форматы МАРК и ONIX. Организация деятельности по созданию метаданных
12. Классификационные языки. Общие понятия классификации.
13. Классификационные языки. УДК.
14. Классификационные языки. ГРНТИ.

15. Проблемы и перспективы применения информационных классификаций в ЭБ
16. Вербальные языки. Общее описание и история развития вербальных языков
17. Лексика и организация лексики в вербальных языках.
18. Информационно-поисковый тезаурус. Принципы создания и ведения тезауруса в УИС «Россия».
19. Грамматика вербальных ИПЯ традиционных АИПС. Методика индексирования.
20. Грамматики вербальных языков современных ЭБ. Организация поиска с использованием вербальных ИПЯ
21. ЛО фактографических и комплексных АИС. Общие понятия фактографии.
22. Интегрированные и комбинированные документально-фактографические системы.
23. Автоматическая обработка текста (АОТ). Виды процессов автоматической обработки текста. Морфологический анализ текста.
24. АОТ. Синтаксический анализ.
25. АОТ. Позиционные методы анализа текста. Суперсинтаксический анализ.
26. АОТ. Семантический анализ. Статистические методы.
27. АОТ. Требования к автоматическому индексированию.
28. Лингвистические банки данных и компьютерная лексикография. Основные типы словарей в АИС. Примеры организации лингвистических банков данных в АИС.
29. Обмен словарями и коммуникативные форматы словарей
30. Лингвистические банки данных в Интернете (+ самостоятельный анализ)
31. Основы компьютерной лексикографии

6. Учебно-методическое и информационное обеспечение дисциплины

6.1. Список источников и литературы

Литература

Основная

1. Волосатова, Т. М. Информатика и лингвистика: учеб. пособие / Т.М. Волосатова, Н.В. Чичварин. — Москва: ИНФРА-М, 2018. — 196 с. URL: https://znanium.com/catalog/product/938009.
2. Тарланов, З. К. Методы лингвистического анализа: для вузов / З. К. Тарланов. — 2-е изд., испр. и доп. — Москва: Издательство Юрайт, 2019. — 236 с.— URL: https://urait.ru/bcode/420842.
3. Казарин, Ю. В. Лингвистический анализ текста: учебное пособие для академического бакалавриата / Ю. В. Казарин; под научной редакцией Л. Г. Бабенко. — 2-е изд. — Москва: Издательство Юрайт, 2019; Екатеринбург: Изд-во Урал. ун-та. — 132 с. — URL: https://urait.ru/bcode/441460.

Дополнительная

1. Астраханцев Н. А. (Институт Системного Программирования РАН). Методы автоматического построения и обогащения неформальных онтологий [Текст] / Н. А. Астраханцев, Д. Ю. Турдаков // Программирование. - 2013. - № 1. - С. 23-35. - Библиогр.: с. 31-35 (70 назв.).
2. Старичкова Ю. В. (аспирант; Национальный исследовательский университет "Высшая школа экономики", Москва). Структурная сложность орграфов: некоторые математические модели и их приложения [Текст] / Ю. В. Старичкова // Научно-техническая информация. Сер. 2, Информационные процессы и системы. - 2013. - № 2. - С. 1-7. - Библиогр.: с. 7 (8 назв.). - Ил.: 7 рис., 2 табл.
3. Яцко В. А. (доктор филологических наук; профессор). Компьютерная лингвистика или лингвистическая информатика?[Текст] / В. А. Яцко // Научно-техническая информация. Сер. 2, Информационные процессы и системы. - 2014. - № 5. - С. 1-10. - Библиогр.: с. 10 (35 назв.). - Ил.: 1 табл.
4. Клышинский Э. С. (кандидат технических наук; доцент). Методика автоматизации проверки полноты технической отчетной документации[Текст] / Э. С. Клышинский, Я. Б. Калачев, В. В. Жаднов // Научно-техническая информация. Сер. 2, Информационные процессы и системы. - 2014. - № 5. - С. 11-15. - Примеч. в сносках. - Библиогр.: с. 14-45 (14 назв.). - Ил.: 2 табл., 2 рис.
5. Ингерсолл Грант С. Обработка неструктурированных текстов. Поиск, организация и манипулирование: Практическое пособие; ВО - Бакалавриат. - Москва: ДМК Пресс, 2015. - 414 с. Ссылка на ресурс: http://new.znanium.com/go.php?id=1027786.
6. Микрин Евгений Анатольевич. Система автоматического анализа текстовых документов // Проблемы управления безопасностью сложных систем. - М.: РГГУ, 2003. - Ч. 1. - С. 197-200.
7. Шаров Ю. Введение в базы данных: Знакомство с компьютером. Обработка текстов. Электронные таблицы. Банки данных. - М.: ABF, 1995. - 383с.: ил. - (Компьютер для носорога; Кн.3.Носорог в море данных). - ISBN 5-87484-015-X: 9345.00.
8. Сафронов И. К. Задачник-практикум по информатике. - Санкт-Петербург: БХВ-Петербург, 2002. - 428 с. - ISBN 978-5-94157-186-0.
9. Сатунина Анна Евгеньевна. E-Learning - курс "Автоматизированная обработка текстов на естественном языке" / Сатунина А. Е., Акатьева М. А., Лазаренко Е. В. // Тенденции и перспективы развития информационных технологий в высшей школе. - М. : РГГУ, 2005. - С. 110-123.

10. Федорец О. В. Хранилище полных текстов для доступа пользователей через электронный каталог поступлений ВИНИТИ / О. В. Федорец, А. М. Фишер, А. А. Батюшко // Научные и технические библиотеки. - 2007. - N 2. - С. 72-78. - Библиогр.: с. 78 (3 назв.).
11. Информационные технологии [Электронный ресурс]: учебник. - 2-е изд., перераб. и доп.-Москва: Издательство "ФОРУМ": Издательский Дом "ИНФРА-М",2008.-608 с. - ISBN 978-5-91134-178-7. Ссылка на ресурс: http://znanium.com/go.php?id=150600.
12. Антонов Александр Викторович. Аналитика и поиск информации в полнотекстовых базах данных / А. В. Антонов // Научно-техническая информация. Сер. 2, Информационные процессы и системы. - 2009. - N 4. - С. 21-28. - Примеч. в сносках. - Библиогр.: с. 28 (4 назв.). - Ил.: 2 рис.
13. Бабкин Э. (канд. техн. наук; проф.; зав. каф. информ. систем и технологий фак. бизнес-информатики ГУ-ВШЭ). Использование онтологий в задачах семантического анализа / Э. Бабкин, А. Шишин // Проблемы теории и практики управления. - 2009. - N 9. - С. 71-79. - Библиогр.: с. 79 (5 назв.).
14. Турдаков Д. Ю. (МГТУ им. М. В. Ломоносова, Факультет вычислительной математики и кибернетики). Автоматическое разрешение лексической многозначности терминов на основе сетей документов / Д. Ю. Турдаков, С. Д. Кузнецов // Программирование. - 2010. - N 1. - С. 16-28. - Библиогр.: с. 27-28 (27 назв.).
15. Колпаков Ю. А. (кандидат технических наук; доцент кафедры информационных технологий и телекоммуникаций Российского государственного торгово-экономического университета (РГТУ), Москва). Проблемы анализа вербальных моделей в экономике[Текст] / Ю. А. Колпаков, Е. В. Романова // Научно-техническая информация. Сер. 2, Информационные процессы и системы. - 2010. - N 11. - С. 26-29. - Библиогр.: с. 29 (8 назв.).

6.2. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

1. Русский национальный корпус, http://www.ruscorpora.ru
2. Частотный словарь Сергея Шарова, http://www.artint.ru/projects/frqlist.asp
3. Поисковые машины и поисковая оптимизация, http://searchengines.ru
4. РОМИП, http://romip.narod.ru
5. Список стоп-слов, http://forum.searchengines.ru/showthread.php?postid=7670
6. Страница Андрея Коваленко, http://linguist.nm.ru
7. Сайт "Автоматическая Обработка Текста", http://www.aot.ru
8. Грамматика русского языка, http://rusgram.narod.ru
9. Библиотека на www.nigma.ru
10. Лингвоанализатор Дмитрия Хмелева http://www.rusf.ru/books/analysis/.
11. Синтаксис языка запросов Яндекса на Livejournal.com http://www.livejournal.com/community/kubok/45852.html
12. Страница Леонида Бойцова http://itman.narod.ru/
13. Морфологический анализатор mystem
14. TextAnalyst http://www.analyst.ru/
15. Учебные материалы на www.informationretrieval.org

6.3. Перечень современных профессиональных баз данных и информационно-справочных систем

№п/п	Наименование
1	Международные реферативные наукометрические БД, доступные в рамках

	национальной подписки в 2021 г. Web of Science Scopus
2	Профессиональные полнотекстовые БД, доступные в рамках национальной подписки в 2021г. Журналы Oxford University Press ProQuest Dissertation & Theses Global SAGE Journals Журналы Taylor and Francis
3	Компьютерные справочные правовые системы Консультант Плюс, Гарант

7. Материально-техническое обеспечение дисциплины

№п/п	Наименование специальных помещений и помещений для самостоятельной работы	Оснащенность специальных помещений и помещений для самостоятельной работы	Перечень лицензионного программного обеспечения. Реквизиты подтверждающего документа		
			Наименование ПО	Лицензия/сертификат/заказ	Дата лицензии
1.	Лаборатория информатики – ауд. № 203	1 компьютер преподавателя, 12 компьютеров обучающихся, маркерная доска, проектор	Windows 7 Microsoft office 2010 Pro Microsoft Visual Professional 2019 Mozilla Firefox 52.8.1 ESR Matlab Mathcad Education - University edition Kaspersky Endpoint Security Платформа ZOOM	68526624 49420326 63202190 свободный доступ 647526 2996385 17E0-181226-094912-873-979 лицензионное	без даты 08.12.2011 без даты свободный доступ без даты 14.06.2019 26.12.2018

8. Обеспечение образовательного процесса для лиц с ограниченными возможностями здоровья

В ходе реализации дисциплины используются следующие дополнительные методы обучения, текущего контроля успеваемости и промежуточной аттестации обучающихся в зависимости от их индивидуальных особенностей:

- для слепых и слабовидящих:

- лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением;
- письменные задания выполняются на компьютере со специализированным программным обеспечением, или могут быть заменены устным ответом;
- обеспечивается индивидуальное равномерное освещение не менее 300 люкс;
- для выполнения задания при необходимости предоставляется увеличивающее устройство; возможно также использование собственных увеличивающих устройств;
- письменные задания оформляются увеличенным шрифтом;
- экзамен и зачёт проводятся в устной форме или выполняются в письменной форме на компьютере.

- для глухих и слабослышащих:

- лекции оформляются в виде электронного документа, либо предоставляется звукоусиливающая аппаратура индивидуального пользования;
- письменные задания выполняются на компьютере в письменной форме;
- экзамен и зачёт проводятся в письменной форме на компьютере; возможно проведение в форме тестирования.

- для лиц с нарушениями опорно-двигательного аппарата:

- лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением;

- письменные задания выполняются на компьютере со специализированным программным обеспечением;

- экзамен и зачёт проводятся в устной форме или выполняются в письменной форме на компьютере.

При необходимости предусматривается увеличение времени для подготовки ответа.

Процедура проведения промежуточной аттестации для обучающихся устанавливается с учётом их индивидуальных психофизических особенностей. Промежуточная аттестация может проводиться в несколько этапов.

При проведении процедуры оценивания результатов обучения предусматривается использование технических средств, необходимых в связи с индивидуальными особенностями обучающихся. Эти средства могут быть предоставлены университетом, или могут использоваться собственные технические средства.

Проведение процедуры оценивания результатов обучения допускается с использованием дистанционных образовательных технологий.

Обеспечивается доступ к информационным и библиографическим ресурсам в сети Интернет для каждого обучающегося в формах, адаптированных к ограничениям их здоровья и восприятия информации:

- для слепых и слабовидящих:

- в печатной форме увеличенным шрифтом;
- в форме электронного документа;
- в форме аудиофайла.

- для глухих и слабослышащих:

- в печатной форме;
- в форме электронного документа.

- для обучающихся с нарушениями опорно-двигательного аппарата:

- в печатной форме;
- в форме электронного документа;
- в форме аудиофайла.

Учебные аудитории для всех видов контактной и самостоятельной работы, научная библиотека и иные помещения для обучения оснащены специальным оборудованием и учебными местами с техническими средствами обучения:

- для слепых и слабовидящих:

- устройством для сканирования и чтения с камерой SARA CE;
- дисплеем Брайля PAC Mate 20;
- принтером Брайля EmBraille ViewPlus;

- для глухих и слабослышащих:

- автоматизированным рабочим местом для людей с нарушением слуха и слабослышащих;
- акустический усилитель и колонки;

- для обучающихся с нарушениями опорно-двигательного аппарата:

- передвижными, регулируемыми эргономическими партами СИ-1;
- компьютерной техникой со специальным программным обеспечением.

9. Методические материалы

9.1. Планы практических занятий и методические указания по их организации и проведению

№ темы	№ п/п	Содержание
Тема 2	1	**Практическая работа 1.** Тема. Обработка текстовой информации. Цель. Изучить правила конструирования строк, основные действия над строками, получить практические навыки решения задач обработки текстов. Содержание работы 1. Изучить правила конструирования строк. 2. Изучить стандартные функции работы со строками. 3. Спроектировать и отладить программу решения задачи, выбрав и обосновав, наиболее удобную для отображения и обработки текста структуру данных.
Тема 3	2	**Практическая работа 2** Тема. Машинный перевод. Сравнить качесто двух-трех систем машинного перевода из списка: ПРОМТ, SYSTRAN, Babel Fish, Free Translation, ЭТАП. Сравните разультаты перевода художественных, технических, газетных тестов, личных электронных писем. Сравните результат работы автомата с переводом, выполненным человеком (например, инструкции, художественные тексты). **Практическая работа 3**. Тема. Поиск в Интернете. Изучите официальное и неофициальное описание языка запросов Яндекса. Изучите основы поиска и операторы языка запросов Google. Пройдите зачет на странице Кубка Яндекса ("тренировочная игра"). Попробуйте пользоваться разными машинами поиска. Оцените субъективно удобство и качество поиска. Чего не хватает? Что лишнее? Распечатайте диплом "Кубка Яндекса". **Практическая работа 4** Тема. Морфологический анализатор. Изучите опции и формат выходной информации морфологического анализатора mystem. Поэкспериментируйте с разными текстами: литературные/технические; русский/английский; Радищев/Пелевин. Как обрабатываются ошибки/опечатки? Как обрабатываются незнакомые слова? Приведите примеры грамматической омонимии. Вычислите степень неоднозначности разбора на небольшом тексте (100-200 слов) - отношение разборы/слова. Предложите методы разрешения неоднозначностей.
Тема 4		**Практическая работа 5**. Цель работы: Тема. Изучение вопросно-ответной системы – поиска в базе знаний по запросу на естественном языке. Соотношение предметов изучения компьютерной лингвистики и управления знаниями на примере вопросно-ответной (QA) системы.

		Программа GEOBASE. **(http://www.pdc.dk/vipexamples/cgiexamples/geobase.htm)** **Практическая работа 6**. Цель работы: Изучение программы грамматического разбора (парсера). Обучение построению программ обработки запроса на ЕЯ. Программа SENAN (http://www.pdc.dk/vipexamples/cgiexamples/sen_an.htm)
Тема 5		**Практическая работа 7**. http://www.manifestation.com/neurotoys/eliza.php3 - поговорите с Элизой, выделите в своем диалоге различные типы реплик Элизы. **Практическая работа 8** Тема. Бинарная классификация с помощью SVM. Ознакомиться с описанием реализации SVM^{light}. Ознакомиться с описанием примера 1. Построить классификатор на основе данных, провести тестирование. Редактировать данные: сократить число положиетельных примеров и т.п. Подготовить данные базы спама в формате SVM. Разделить данные на обучающее и тестовое множество. Построить и протестировать классификатор.

9.2. Методические рекомендации по подготовке письменных работ.

Методические рекомендации по подготовке письменных работ, требования к их содержанию и оформлению

Порядок составления и оформления отчета о практической работе

В значительной мере эффективность решения задачи по выполнению практической работы зависит от качества соответствующего отчета. Для этого необходимо соблюдать следующие основные требования по составлению и оформлению отчета, обусловленные соответствующими нормативными документами. Текст отчета должен быть лаконичным и вместе с тем информативным. Текст должен быть изложен с соблюдением правил грамматики. Отчет составляется с обязательным составлением следующих разделов:

1. Заголовок отчета.
2. Цели работы.
3. Методика работы.
4. Порядок выполнения работы (этапы работы).
5. Выводы по работе.

1. В **заголовке отчета** приводятся наименования идентифицирующих признаков: **Отчет о практической работе № 1** по теме, например, «**Обработка текстовой информации**», ниже указываются данные студента (фамилия и инициалы, вид обучения, специальность, курс, группа).

2. В разделе **Цель работы** формулируется цели работы студента в соответствии с содержанием раздела «Постановка задачи» данной работы и индивидуального задания студенту на работу.

3. В разделе **Методика работы указывается** методика работы в соответствии с имеющейся формулировкой в разделе «Методика работы» данной работы и при необходимости уточняется в зависимости от содержания конкретного варианта задания студенту на практическую работу.

4. **Порядок выполнения работы**. Приводятся номера и наименования этапов работы, предусмотренные для работы данного Практикума. По каждому из этапов приводится описание выполненных студентом работ, направленных на достижение цели работы. Пропуск какого-либо из этапов работы Практикума не допускается. В рамках этапов помещается соответствующий иллюстративный материал - таблицы, рисунки (графики), полученные по ходу решения задачи работы. Обозначение иллюстративного материала выполняется в соответствии с правилами, принятыми для публикаций. Обозначение каждой таблицы и рисунка должно иметь номер и наименование. Внутри каждого отчета таблицы и рисунки обозначаются соответственно сквозными номерами. Обозначение таблицы указывается над таблицей, а обозначение рисунка под рисунком. Приводимые в тексте данной работы примеры включать в отчет не разрешается. Применяется только материал, полученный в ходе работы студентом по соответствующему заданию, полученному от преподавателя.

5. Последним разделом отчета являются **выводы** по работе. Это самая сложная и трудная часть работы. Очень важно, чтобы выводы отражали методику, технологию, применяемые программно-аппаратные средства решения задачи. Полезно каждому из этапов работы формулировать не менее одного вывода. Вывод может содержать от одного до трех предложений. Формулировки выводов должны быть конкретными, информативными, лаконичными, по возможности подкрепляться количественными данными.

Оформление отчета выполняется с учетом общепринятых правил. Графическая часть отчетов должна соответствовать правилам графического оформления. Текст отчета набирается в редакторе Word через 1,5 интервала, 14 кегль. Следует использовать шрифт Times New Roman. Заголовки разделов и подразделов выделяются жирным шрифтом. После окончания оформления отчета он проверяется студентом на предмет качество содержания и формы. При условии обнаружения ошибок последние исправляются. После устранения дефектов отчета его экранная форма, или принтерная распечатка предъявляется преподавателю. При условии обнаружения преподавателем ошибок в отчете студент их исправляет и предъявляет отчет преподавателю повторно. Если ошибок нет, то отчет принимается и сохраняется на жестком диске.

Отчет по работе сохраняется студентом в виде отдельного файла. В имени файла указывается фамилия студента и номер выполненной работы. Файл сохраняется в папке с фамилией студента в папке соответствующей студенческой группы. Папка группы создается на первом занятии. В имени папки группы должен присутствовать индекс группы. Папка группы включается в папку «Мои документы».

Приложение 1

АННОТАЦИЯ ДИСЦИПЛИНЫ

Дисциплина реализуется на факультете информационных систем и безопасности ИИНТБ РГГУ, кафедрой информационных технологий и систем.

***Цель* дисциплины:** формирование у студентов научного подхода к освоению, созданию и использованию в гуманитарной сфере интеллектуальных информационных систем, основанных на текстовых базах знаний и естественно-языковых средствах коммуникации.

Задачи: освоить общие принципы построения систем обработки текстов; раскрыть структуру лингвистических процессоров и модульный принцип их построения; освоить методы построения модулей лингвистических процессоров (графематического, морфологического, синтаксического); ознакомиться с принципами и методами построения модуля семантического анализа; освоить основы компьютерной лексикографии; дать представление о возможностях автоматического создания текстов.

Дисциплина направлена на формирование следующих компетенций:

ПК-3 Способен проектировать информационные системы по видам обеспечения

В результате освоения дисциплины обучающийся должен:

знать: *общие* принципы построения систем автоматической обработки текстов (письменного и устного); структуру систем синтеза и анализа; этапы и уровни автоматической обработки текста; построение графематического, морфологического и синтаксического анализа; практические возможности современных систем автоматической обработки естественного языка

уметь: работать с автоматическими словарями; охарактеризовать принципы морфологического, синтаксического анализа; осуществлять реферирование текста с использованием компьютерных средств.

обладать навыками работы с инструментами автоматической обработки естественного языка.

Рабочей программой предусмотрены следующие виды контроля: текущий контроль успеваемости в форме защиты отчетов по практическим работам, промежуточная аттестация в форме зачета.

Общая трудоёмкость дисциплины составляет 3 зачетные единицы.