Минобрнауки России

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

(ФГБОУ ВО «ВГУ»)

УТВЕРЖДАЮ

Заведующий кафедрой

Каменский Михаил Игоревич

Кафедра функционального анализа и операторных уравнений

20.03.2025г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

Б1.О.13.03 Автоматизированная обработка текстов

1. Код и наименование направления подготовки/специальности:

10.05.04 Информационно-аналитические системы безопасности

2. Профиль подготовки/специализация:

Автоматизация информационно-аналитической деятельности

3. Квалификация (степень) выпускника:

Специалитет

4. Форма обучения:

Очная

5. Кафедра, отвечающая за реализацию дисциплины: Кафедра функционального анализа и операторных уравнений

6. Составители программы:

Копытин Алексей Вячеславович, доцент кафедры функционального анализа и операторных уравнений

7. Рекомендована:

НМС математического факультета, протокол № 0500-03 от 18.03.2025г.

8. Учебный год:

2029-2030

9.Цели и задачи учебной дисциплины:

Цели изучения дисциплины:

- освоение основ автоматизированной обработки текстов (АОТ), овладение базовыми навыками работы с существующими программными средствами АОТ и лингвистическими ресурсами. Знание принципов компьютерной обработки текстов и умение применять автоматизированные технологии обработки текстов при решении прикладных информационно-аналитических задач.

Задачи учебной дисциплины:

- изучение основных видов автоматизированных систем обработки и анализа естественно-языковых текстов.

-изучение основных математических моделей, методов и алгоритмов решения типовых задач обработки и анализа естественно-языковых текстов в ИАС.

10. Место учебной дисциплины в структуре ООП:

дисциплина Автоматизированная обработка текстов относится к обязательной части блока Б1.

11. Планируемые результаты обучения по дисциплине/модулю (знания, умения, навыки),соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями выпускников) и индикаторами их достижения:

Код и название компетенции	Код и название индикатора компетенции	Знания, умения, навыки
ОПК-1.1. Способен разрабатывать и применять автоматизированные технологии обработки естественно-языковых текстов;	ОПК-1.1. .1 знает основные типы задач обработки и анализа естественно-языковых текстов, основные типы автоматизированной информационно-аналитической поддержки процессов принятия решений	Знать: основные типы задач обработки и анализа естественно-языковых текстов, основные типы автоматизированной информационно-аналитической поддержки процессов принятия решений. Уметь: решать основные типы задач обработки и анализа естественно-языковых текстов, основные типы автоматизированной информационно-аналитической поддержки процессов принятия решений. Владеть: основными типами задач обработки и анализа естественно-языковых текстов, основными типами автоматизированной информационно-аналитической поддержки процессов принятия решений.

ОПК-1.1. Способен разрабатывать и применять автоматизированные технологии обработки естественно-языковых текстов;	ОПК-1.1. .2 знает основные виды автоматизированных систем обработки и анализа естественно-языковых текстов	Знать: основные типы задач обработки и анализа естественно-языковых текстов, основные типы автоматизированной информационно-аналитической поддержки процессов принятия решений. Уметь: решать основные типы задач обработки и анализа естественно-языковых текстов, основные типы автоматизированной информационно-аналитической поддержки процессов принятия решений. Владеть: основными типами задач обработки и анализа естественно-языковых текстов, основными типами автоматизированной информационно-аналитической поддержки процессов принятия решений.

Код и название компетенции	Код и название индикатора компетенции	Знания, умения, навыки

ОПК-1.1. Способен разрабатывать и применять автоматизированные технологии обработки естественно-языковых текстов;	ОПК-1.1. .3 знает основные математические модели, методы и алгоритмы решения типовых задач обработки и анализа естественно-языковых текстов в ИАС	Знать: основные типы задач обработки и анализа естественно-языковых текстов, основные типы автоматизированной информационно-аналитической поддержки процессов принятия решений. Уметь: решать основные типы задач обработки и анализа естественно-языковых текстов, основные типы автоматизированной информационно-аналитической поддержки процессов принятия решений. Владеть: основными типами задач обработки и анализа естественно-языковых текстов, основными типами автоматизированной информационно-аналитической поддержки процессов принятия решений.

ОПК-1.1. Способен разрабатывать и применять автоматизированные технологии обработки естественно-языковых текстов;	ОПК-1.1. .4 умеет проводить оценку качества и осуществлять выбор автоматизированной технологии семантической обработки текстов в конкретных условиях решения прикладных информационно-аналитических задач	Знать: основные типы задач обработки и анализа естественно-языковых текстов, основные типы автоматизированной информационно-аналитической поддержки процессов принятия решений. Уметь: решать основные типы задач обработки и анализа естественно-языковых текстов, основные типы автоматизированной информационно-аналитической поддержки процессов принятия решений. Владеть: основными типами задач обработки и анализа естественно-языковых текстов, основными типами автоматизированной информационно-аналитической поддержки процессов принятия решений.

Код и название компетенции	Код и название индикатора компетенции	Знания, умения, навыки
ОПК-1.1. Способен разрабатывать и применять автоматизированные технологии обработки естественно-языковых текстов;	ОПК-1.1. .5 умеет применять автоматизированные технологии семантической обработки текстов при решении прикладных информационно-аналитических задач, в том числе для автоматизированной информационно-аналитической поддержки процессов принятия решений	Знать: основные типы задач обработки и анализа естественно-языковых текстов, основные типы автоматизированной информационно-аналитической поддержки процессов принятия решений. Уметь: решать основные типы задач обработки и анализа естественно-языковых текстов, основные типы автоматизированной информационно-аналитической поддержки процессов принятия решений. Владеть: основными типами задач обработки и анализа естественно-языковых текстов, основными типами автоматизированной информационно-аналитической поддержки процессов принятия решений.

ОПК-1.1. Способен разрабатывать и применять автоматизированные технологии обработки естественно-языковых текстов;	ОПК-1.1. .6 владеет навыками работы с программными системами, реализующими автоматизированные технологии семантической обработки текстов	Знать: основные типы задач обработки и анализа естественно-языковых текстов, основные типы автоматизированной информационно-аналитической поддержки процессов принятия решений. Уметь: решать основные типы задач обработки и анализа естественно-языковых текстов, основные типы автоматизированной информационно-аналитической поддержки процессов принятия решений. Владеть: основными типами задач обработки и анализа естественно-языковых текстов, основными типами автоматизированной информационно-аналитической поддержки процессов принятия решений.

12. Объем дисциплины в зачетных единицах/час:

4/144

Форма промежуточной аттестации:

Экзамен

13. Трудоемкость по видам учебной работы

Вид учебной работы	Семестр 9	Всего
Аудиторные занятия	72	72
Лекционные занятия	36	36
Практические занятия		0
Лабораторные занятия	36	36
Самостоятельная работа	36	36
Курсовая работа		0
Промежуточная аттестация	36	36
Часы на контроль	36	36
Всего	144	144

13.1. Содержание дисциплины

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
1	Лекции		
1.1	Введение	Автоматизированная обработка текстов на естественном языке (ЕЯ): основные задачи и особенности направления, связь со смежными научными дисциплинами. Естественный язык как сложная система языковых знаков. Уровни языковой системы. Феномены ЕЯ: полисемия, синонимия, омонимия. Лингвистические процессоры и лингвистические ресурсы. Этапы анализа текста. Обзор основных приложений АОТ (машинный перевод, информационный поиск, классификация текстов, реферирование и аннотирование, извлечение информации и знаний, анализ тональности, автоматизация редактирования текстов).	https://edu.vsu.ru/course/view.php?id=29818

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК

1.2	Начальные этапы обработки текста	Графематический анализ и сегментация текста. Токенизация и разбиение на предложения. Виды токенов, обработка сложных случаев. Основные понятия морфологии: словоформа, морфема, аффикс, корень, основа, флексия. Словоизменительная парадигма и морфологические параметры. Словарные и бессловарные модели морфологии. Автоматический морфологический анализ и синтез. Виды морфологического анализа: стемминг, лемматизация, полный морфоанализ. Принципы морфоанализа на базе словаря основ или словаря словоформ.	https://edu.vsu.ru/course/view.php?id=29818
		Морфологические процессоры для русского языка.	

1.3	Статистические характеристики текстов и корпусная лингвистика	Статистика словоупотреблений в текстах на ЕЯ. Абсолютные и относительные частоты словоформ и лексем. Закон Ципфа-Мандельброта и его интерпретация. Принцип Парето. Формула Кондона. Соотношение длины слова и его частоты. Глоттохронология. Статистика встречаемости символов и буквосочетаний: биграмм и триграмм, N-грамм. Задачи АОТ, решаемые на базе статистики символов. Задачи корпусной лингвистики. Коллекции и корпуса текстов. Характеристики и состав типичного корпуса. Обзор корпусов. Национальный корпус русского языка. Статистика N-грамм для слов. Понятие статистической языковой модели. Применение статистической (вероятностной) модели для разрешения морфологической омонимии. Использование статистики для автоматического выделения устойчивых словосочетаний языка. Меры устойчивости сочетаний.	https://edu.vsu.ru/course/view.php?id=29818

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК

1.4	Подходы к автоматическому анализу синтаксиса и семантики текста	Задачи синтаксического анализа ЕЯ. Синтаксические деревья непосредственных составляющих и деревья зависимостей. Синтаксические связи слов. Понятия синтаксического предиката и модели управления. Синтаксический разбор на базе контекстносвободных грамматик. Примеры синтаксических парсеров. Частичный синтаксический анализ. Понятие синтаксической сегментации текста. Автоматическое выделение словосочетаний (именных, предложных групп). Основные способы представления смысла текста и модели представления знаний в искусственном интеллекте: семантические сети, язык предикатов. Семантический анализ текста на основе семантико-синтаксических моделей управления. Связный текст (дискурс), его особенности. Смысловая и синтаксическая связность. Анафорические ссылки, лексические повторы, дискурсивные слова. Сверхфразовые единства. Композиционные и дискурсивные особенности текстов, их учет в задачах АОТ.	https://edu.vsu.ru/course/view.php?id=29818
1.5	Лингвистические ресурсы	Словари для автоматической обработки текстов. Виды словарей. Терминологические словари и автоматизация их построения. Смысловые (парадигматические) отношения лексических единиц. Синонимия и лексическая многозначность. Тезаурус как словарь с семантическими связями единиц. Информационно-поисковые тезаурусы и рубрикаторы, их применение в АОТ. Понятие онтологии. Лингвистические онтологии. Лексические ресурсы WordNet и EvroNet. Методологии создания онтологий.	https://edu.vsu.ru/course/view.php?id=29818

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
1.6	Прикладные задачи АОТ	Подходы к разработке приложений АОТ: инженерный подход, основанный на лингвистических правилах, и подход, основанный на машинном обучении. Основные показатели качества работы систем АОТ: точность, полнота, Fмера. Информационный поиск в массивах полнотекстовых документов: основные понятия. Индексирование текстов для информационного поиска. Векторная модель документа. Булевский поиск, ранжированный поиск. Оценка релевантности документа. Поиск в сети Интернет, принципы работы поисковых машин. Классификация текстов как типичная задача обработки текстов в области TextMining. Обзор методов машинной классификации. Выбор признаков и метрик. Особенности кластеризации текстов. Рубрицирование текстовых документов. Обзор задач АОТ, решаемых на основе классификации текстов. Автоматическое реферирование и аннотирование документов как смежные задачи информационного поиска. Основные стратегии сжатия текста. Типы аннотаций. Обзорное реферирование. Оценка качества аннотаций. Машинный перевод. Стратегии машинногоперевода, основанного на лингвистических правилах. Статистический машинный перевод: особенности и виды. Принципы создания статистического переводчика. Извлечение информации и знаний из текстов: особенности задачи и типы извлекаемых объектов. Понятие лингвистического шаблона для	https://edu.vsu.ru/course/view.php?id=29818

		извлечения информации. Инструментальные программные средства для построения систем извлечения информации из текстов. Извлечение знаний под управлением онтологий в системах класса OntosMiner. Автоматический анализ тональности текстов и извлечение мнений из текстов: особенности и подходы к решению. Анализ тональности как задача классификации.	
2	Практические занятия		

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
2.1	нет		
3	Лабораторные занятия		
3.1	Введение	Автоматизированная обработка текстов на естественном языке (ЕЯ): основные задачи и особенности направления, связь со смежными научными дисциплинами. Естественный язык как сложная система языковых знаков. Уровни языковой системы. Феномены ЕЯ: полисемия, синонимия, омонимия. Лингвистические процессоры и лингвистические ресурсы. Этапы анализа текста. Обзор основных приложений АОТ (машинный перевод, информационный поиск, классификация текстов, реферирование и аннотирование, извлечение информации и знаний, анализ тональности, автоматизация редактирования текстов).	https://edu.vsu.ru/course/view.php?id=29818

3.2	Начальные этапы обработки текста	Графематический анализ и сегментация текста. Токенизация и разбиение на предложения. Виды токенов, обработка сложных случаев. Основные понятия морфологии: словоформа, морфема, аффикс, корень, основа, флексия. Словоизменительная парадигма и морфологические параметры. Словарные и бессловарные модели морфологии. Автоматический морфологический анализ и синтез. Виды морфологического анализа: стемминг, лемматизация, полный морфоанализ. Принципы морфоанализа на базе словаря основ или словаря словоформ. Морфологические процессоры для русского языка.	https://edu.vsu.ru/course/view.php?id=29818

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК

3.3	Статистические характеристики текстов и корпусная лингвистика	Статистика словоупотреблений в текстах на ЕЯ. Абсолютные и относительные частоты словоформ и лексем. Закон Ципфа-Мандельброта и его интерпретация. Принцип Парето. Формула Кондона. Соотношение длины слова и его частоты. Глоттохронология. Статистика встречаемости символов и буквосочетаний: биграмм и триграмм, N-грамм. Задачи АОТ, решаемые на базе статистики символов. Задачи корпусной лингвистики. Коллекции и корпуса текстов. Характеристики и состав типичного корпуса. Обзор корпусов. Национальный корпус русского языка. Статистика N-грамм для слов. Понятие статистической языковой модели. Применение статистической (вероятностной) модели для разрешения морфологической омонимии. Использование статистики для автоматического выделения устойчивых словосочетаний языка. Меры устойчивости сочетаний.	https://edu.vsu.ru/course/view.php?id=29818
3.4	Подходы к автоматическому анализу синтаксиса и семантики текста	Задачи синтаксического анализа ЕЯ. Синтаксические деревья непосредственных составляющих и деревья зависимостей. Синтаксические связи слов. Понятия синтаксического предиката и модели управления. Синтаксический разбор на базе контекстносвободных грамматик. Примеры синтаксических парсеров. Частичный синтаксический анализ. Понятие синтаксической сегментации текста. Автоматическое выделение словосочетаний (именных, предложных групп). Основные способы представления смысла текста и модели представления знаний в искусственном интеллекте: семантические сети, язык предикатов. Семантический анализ текста на основе семантико-синтаксических моделей управления. Связный текст (дискурс), его особенности. Смысловая и синтаксическая связность. Анафорические ссылки, лексические повторы, дискурсивные слова.	https://edu.vsu.ru/course/view.php?id=29818

		Сверхфразовые единства. Композиционные и дискурсивные особенности текстов, их учет в задачах АОТ.	

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
3.5	Лингвистические ресурсы	Словари для автоматической обработки текстов. Виды словарей. Терминологические словари и автоматизация их построения. Смысловые (парадигматические) отношения лексических единиц. Синонимия и лексическая многозначность. Тезаурус как словарь с семантическими связями единиц. Информационно-поисковые тезаурусы и рубрикаторы, их применение в АОТ. Понятие онтологии. Лингвистические онтологии. Лексические ресурсы WordNet и EvroNet. Методологии создания онтологий.	https://edu.vsu.ru/course/view.php?id=29818

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК

3.6	Прикладные задачи АОТ	Подходы к разработке приложений АОТ: инженерный подход, основанный на лингвистических правилах, и подход, основанный на машинном обучении. Основные показатели качества работы систем АОТ: точность, полнота, Fмера. Информационный поиск в массивах полнотекстовых документов: основные понятия. Индексирование текстов для информационного поиска. Векторная модель документа. Булевский поиск, ранжированный поиск. Оценка релевантности документа. Поиск в сети Интернет, принципы работы поисковых машин. Классификация текстов как типичная задача обработки текстов в области TextMining. Обзор методов машинной классификации. Выбор признаков и метрик. Особенности кластеризации текстов. Рубрицирование текстовых документов. Обзор задач АОТ, решаемых на основе классификации текстов. Автоматическое реферирование и аннотирование документов как смежные задачи информационного поиска. Основные стратегии сжатия текста. Типы аннотаций. Обзорное реферирование. Оценка качества аннотаций. Машинный перевод. Стратегии машинногоперевода, основанного на лингвистических правилах. Статистический машинный перевод: особенности и виды. Принципы создания статистического переводчика. Извлечение информации и знаний из текстов: особенности задачи и типы извлекаемых объектов. Понятие лингвистического шаблона для извлечения информации. Инструментальные программные средства для построения систем извлечения информации из текстов. Извлечение знаний под управлением онтологий в системах класса OntosMiner. Автоматический анализ тональности текстов и извлечение мнений из текстов: особенности и подходы к решению.	https://edu.vsu.ru/course/view.php?id=29818

		Анализ тональности как задача классификации.	

13.2.Темы (разделы) дисциплины и виды занятий

№ п/п	Наименование темы (раздела)	Лекционные занятия	Практические занятия	Лабораторные занятия	Самостоятельная работа	Всего

1	Введение	6		3	6	15
2	Начальные этапы обработки текста	5		5	6	16
3	Статистические характеристики текстов и корпусная лингвистика	6		6	6	18
4	Подходы к автоматическому анализу синтаксиса и семантики текста	5		6	6	17
5	Лингвистические ресурсы	4		4	6	14
6	Прикладные задачи АОТ	10		12	6	28
		36	0	36	36	108

14. Методические указания для обучающихся по освоению дисциплины

Рекомендуется работа с конспектами лекций, презентационным материалом, выполнение всех лабораторных и контрольных работ.

15. Перечень основной и дополнительной литературы, ресурсов интернет, необходимых для освоения дисциплины

№ п/п	Источник

1	Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большакова Е.И. и др. – М.: Изд-во НИУ ВШЭ, 2017. — 269 с.
2	Васильев В.Г., Кривенко М.П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008.

б) дополнительная литература:

№ п/п	Источник
1	Апресян Ю.Д. и др. Лингвистическое обеспечение системы ЭТАП-2. М.: Наука, 1989.
№ п/п	Источник
2	Барсегян А.А. и др. Технологии анализа данных: DataMining, VisualMining, TextMining, OLAP–2-еизд. –СПб.: БХВ-Петербург, 2008.
3	Болховитянов А.В., Гусев А.В., Чеповский А.М. Морфологические модели компьютерной лингвистики: учеб. пособие –М. МГУП, 2010.

в) информационные электронно-образовательные ресурсы:

№ п/п	Источник

16. Перечень учебно-методического обеспечения для самостоятельной работы

№ п/п	Источник
1	Кобозева И.М. Лингвистическая семантика. –М., 2009.
2	Чатуев М.Б., Чеповский А.М. Частотные методы в компьютерной лингвистике: учеб. пособие –М. МГУП, 2011.

17. Образовательные технологии, используемые при реализации учебной дисциплины,включая дистанционные образовательные технологии (ДОТ), электронное обучение (ЭО), смешанное обучение):

В преподавании данной дисциплины сочетаются:

лекции в форме презентаций (которые затем высылаются студентам для их самостоятельной работы); практические занятия по изучению интернет-ресурсов по АОТ, а также прикладного программного обеспечения; домашние задания на применение изученных ресурсов и методов автоматической обработки текстов.

18. Материально-техническое обеспечение дисциплины:

Для лекций и лабораторных занятий по темам дисциплины используется проектор и компьютеры с выходом в сеть Интернет.

19. Оценочные средства для проведения текущей и промежуточной аттестаций

Порядок оценки освоения обучающимися учебного материала определяется содержанием следующих разделов дисциплины:

№ п/п	Разделы дисциплины (модули)	Код компетенции	Код индикатора	Оценочные средства для текущей аттестации
1	Введение	ОПК-1.1.	ОПК-1.1. .1	Практическое задание из пункта 20.1
2	Начальные этапы обработки текста	ОПК-1.1.	ОПК-1.1. .2	Практическое задание из пункта 20.1
3	Статистические характеристики текстов и корпусная лингвистика	ОПК-1.1.	ОПК-1.1. .3	Практическое задание из пункта 20.1
4	Подходы к автоматическому анализу синтаксиса и семантики текста	ОПК-1.1.	ОПК-1.1. .4	Практическое задание из пункта 20.1
5	Лингвистические ресурсы	ОПК-1.1.	ОПК-1.1. .5	Практическое задание из пункта 20.1
6	Прикладные задачи АОТ	ОПК-1.1.	ОПК-1.1. .6	Практическое задание из пункта 20.1

Промежуточная аттестация

Форма контроля - Экзамен

Оценочные средства для промежуточной аттестации

Для оценивания результатов обучения на экзамене используются следующие содержательные показатели (формулируется с учетом конкретных требований дисциплины):

1) знание теоретических основ учебного материала, основных определений, понятий и используемой терминологии;
2) умение проводить обоснование и представление основных теоретических ипрактических результатов (теорем, алгоритмов, методик) с использованием математических выкладок, блок-схем, структурных схем и стандартных описаний к ним;
3) умение связывать теорию с практикой, иллюстрировать ответ примерами, в том числе, собственными, умение выявлять и анализировать основные закономерности, полученные, в том числе, в ходе выполнения лабораторно-практических заданий;
4) умение обосновывать свои суждения и профессиональную позицию по излагаемому вопросу;
5) владение навыками программирования и экспериментирования в рамках выполняемых лабораторных заданий;

Различные комбинации перечисленных показателей определяют критерии оценивания результатов обучения (сформированности компетенций) на зачете:

– высокий (углубленный) уровень сформированности компетенций; – повышенный (продвинутый) уровень сформированности компетенций; – пороговый (базовый) уровень сформированности компетенций.

20 Типовые оценочные средства и методические материалы, определяющие процедурыоценивания

20.1 Текущий контроль успеваемости Примеры

заданий:

1. Реализовать на одном из языков программирования собственный графематический анализатор для русскоязычных текстов и протестировать его на реальных текстах.

2. Провести сравнительный анализ функциональных возможностей двух морфопроцессоров для русского языка.

3. Выполнить статистический анализ двух текстов на русском языке, программно вычислив 5-7 его статистических характеристик (общестатистических, морфологических, лексических) на базе предварительного морфологического анализа слов текста.

4. Для заданного слова русского языка исследовать временные изменения частоты его употребления слова и смысла, рассмотрев его значения и толкования в различных толковых словарях, в Национальном корпусе русского языка, а также в яндекс-новостях.

20.2 Промежуточная аттестация Примеры

вопросов на экзамене:

1. Определите, есть ли в предложении омонимичные словоформы, и если есть, укажите для одной из них все варианты леммы и морфологические характеристики: *Пила – инструмент со множеством резцов*.
2. Покажите на примере, чем словоупотребление отличается от словоформы.
3. Что такое N-грамма? Перечислите все символьные триграммы в словосочетании *на бал*.
4. Объясните понятие валентности слова-предиката. Приведите примеры трех слов предикатов разных частей речи с указанием для них валентностей.
5. На примере 2-5 слов русского языка покажите возможные виды связей лексем в лингвистических онтологиях.
6. Охарактеризуйте модель булевского поиска в массиве документов.

7. Поясните смысл показателей idf и tf.idf.
8. Сравните два основных подхода к машинный переводу (на основе лингвистических правил, статистический перевод).
9. Укажите основные этапы обработки текста при извлечении информации в подходе, основанном на правилах.
10. Назовите и кратко охарактеризуйте две задачи АОТ, которые можно решать с помощью информации о частотах употребления слов.

20.3 Фонд оценочных средств сформированности компетенций студентов, рекомендуемый для проведения диагностических работ

1) закрытые задания (тестовые, средний уровень сложности):

1. Разное количество информации в одном и том же сообщении для разных людей зависит не от...

a) накопленных ими знаний;

b) уровня понимания сообщения;

c) их интереса к сообщению;

d) их уровня владения компьютерной техникой.

Ответ: d

2.Машинный синтаксис — это ...

a) правила строения имен;

b) правила построения слов в более сложные структуры;

c) соотношение слова и его значения;

d) правила перевода письменного символа в устный.

Ответ: a

3.Какие из следующих приложений не являются текстовыми редакторами?

a) MS Excel;

b) Corel WordPerfect;

c) MS Works;

d) Adobe InCopy.

Ответ: a

4. Microsoft Word не включает...

a) функции настольных издательских систем;

b) функцию удалённого доступа;

c) функцию редактирования графических объектов;

d) шаблоны типовых таблиц.

Ответ: b

5. К устройствам ввода данных не относится

a) сканер;

b) принтер;

c) клавиатура;

d) цифровой фотоаппарат.

Ответ: b

2) открытые задания (тестовые, повышенный уровень сложности):

1. Вставьте слово. …… язык — это знаковая система, стихийно возникшая и закрепившаяся в обществе.

Ответ: Естественный

2. Вставьте три слова. OCR — это система ……;

Ответ: автоматического распознавания символов

3. Вставьте два слова. Слово, относящееся к основному содержанию текста и повторяющееся в нем несколько раз, в автоматическом реферировании называется ...

Ответ: ключевым словом.

4. Вставьте два слова. …….— это словарь в специальном машинном формате, предназначенный для применения на ЭВМ пользователем или компьютерной программой.

Ответ: Электронный словарь

5. Вставьте два слова. …….. — это программно-методический комплекс, позволяющий самостоятельно освоить учебный курс или его большой раздел.

Ответ: Компьютерный учебник

Критерии и шкалы оценивания заданий ФОС:

1) Задания закрытого типа (выбор одного варианта ответа, верно/неверно):

- 1 балл – указан верный ответ;
- 0 баллов – указан неверный ответ.

2) Задания закрытого типа (множественный выбор):

- 2 балла – указаны все верные ответы;
- 0 баллов — указан хотя бы один неверный ответ.

3) Задания закрытого типа (на соответствие):

- 2 балла – все соответствия определены верно;
- 0 баллов – хотя бы одно сопоставление определено неверно.

4) Задания открытого типа (короткий текст):

- 2 балла – указан верный ответ;
- 0 баллов – указан неверный ответ.

5) Задания открытого типа (число):

- 2 балла – указан верный ответ;

0 баллов – указан неверный ответ.

Задания раздела 20.3 рекомендуются к использованию при проведении диагностических работ с целью оценки остаточных результатов освоения данной дисциплины (знаний, умений, навыков).