

МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «ВГУ»)

УТВЕРЖДАЮ

Заведующий кафедрой
теоретической и прикладной лингвистики
Шилихина К.М.

30.05.2025 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
Б1.В.02 Введение в компьютерную лингвистику

1. Код и наименование направления подготовки/специальности:

10.05.04 Информационно-аналитические системы безопасности

2. Профиль подготовки/специализация: Автоматизация информационно-аналитической деятельности

3. Квалификация выпускника: специалист по защите информации

4. Форма обучения: очная

5. Кафедра, отвечающая за реализацию дисциплины: кафедра теоретической и прикладной лингвистики

6. Составители программы: Дониная Ольга Валерьевна, кандидат филол. наук, доцент кафедры теоретической и прикладной лингвистики

7. Рекомендована: НМС факультета РГФ, протокол № 8 от 18 апреля 2025 г.

8. Учебный год: 2025/2026

Семестр(ы): 2

9. Цели и задачи учебной дисциплины

Целью освоения учебной дисциплины является ознакомление студентов с основными направлениями, а также с основными прикладными практическими задачами современной компьютерной лингвистики, с формальными, программно-реализуемыми подходами к изучению структур и закономерностей естественных языков, с возможностями применения знаний о языке в новых информационных технологиях. В рамках курса предусматривается ознакомление учащихся с современными программными средствами для решения базовых лингвистических прикладных задач.

Задачи учебной дисциплины:

- показать связь между теоретическим и прикладным лингвистическим знанием,
- сформировать у студентов терминологическую базу и навыки использования лингвистически ориентированных программных продуктов для решения практических задач, связанных с обработкой естественного языка,
- познакомить обучающихся с интеллектуальными системами в лингвистической сфере с целью обучения применения таких систем для решения прикладных задач.

10. Место учебной дисциплины в структуре ООП: Блок Б1. Дисциплины (модули). Часть, формируемая участниками образовательных отношений (вариативная).

Для успешного освоения дисциплины требуются базовые навыки работы с компьютером.

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

Код	Название компетенции	Код(ы)	Индикатор(ы)	Планируемые результаты обучения
ПК-3	Способен решать типовые задачи обработки и анализа информации в информационно-аналитических системах государственных органов, обеспечивающих национальную безопасность	ПК-3.1	Владеет способами решения типовых задач обработки и анализа информации в информационно-аналитических системах	знать: • информационно-лингвистические технологии, технологии автоматической обработки естественного языка и искусственного интеллекта • принципы работы лингвистически ориентированных программных продуктов • основные типы задач обработки и анализа естественно-языковых текстов, • основные виды автоматизированных систем обработки и анализа естественно-языковых текстов уметь: • использовать лингвистически-ориентированные программные системы • подбирать информационно-коммуникационные технологии для наиболее эффективного решения профессиональных задач; • применять информационно-

				лингвистические технологии, технологии автоматической обработки естественного языка и искусственного интеллекта в соответствии с решаемой профессиональной задачей • проводить оценку качества и осуществлять выбор автоматизированной технологии обработки текстов в конкретных условиях решения прикладных информационно-аналитических задач • применять автоматизированные технологии обработки текстов при решении прикладных информационно-аналитических задач, владеть: • навыками работы с программными системами, реализующими автоматизированные технологии автоматической обработки естественного языка
--	--	--	--	--

12. Объем дисциплины в зачетных единицах/час. (в соответствии с учебным планом) — 4 з.е. / 144 ч.

Форма промежуточной аттестации: зачет с оценкой.

13. Трудоемкость по видам учебной работы

Вид учебной работы		Трудоемкость			
		Всего	По семестрам		
			2 семестр	№ семестра	...
Контактная работа		68	68		
в том числе:	лекции	34	34		
	практические	-	-		
	лабораторные	34	34		
	курсовая работа	-	-		
	др. виды(при наличии)				
Самостоятельная работа		76	76		
Промежуточная аттестация (для экзамена)		-	-		
Итого:		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	Прикладная лингвистика как отрасль научного знания	Соотношение теоретической и прикладной лингвистики. Объект и методы исследования в прикладной лингвистике. История развития прикладного лингвистического знания. Современная трактовка термина «прикладная лингвистика». Прикладная лингвистика как одно из направлений Digital Humanities и искусственного интеллекта. Основные направления прикладной лингвистики. Методы исследования в прикладной лингвистике. Основные понятия теории знаний (виды знаний, структуры представления знаний; модель мира). Разработка моделей коммуникации. Гипертекстовые технологии представления текста	
1.2	Компьютерная лингвистика как раздел прикладной лингвистики	Компьютерная лингвистика как междисциплинарное направление исследований. Задачи компьютерной лингвистики. Сложности моделирования естественного языка. Общие этапы и модули обработки текстов. Базовый терминологический аппарат компьютерной лингвистики. Основные направления компьютерной лингвистики. Подходы к построению модулей и систем компьютерной лингвистики. Признаковая модель текста (в том числе BOW – bag of words). Статистическая языковая модель. Методы выделения устойчивых словосочетаний.	
1.3	Системы автоматической обработки звучащей речи. Диалоговые системы и чат-боты	Общие проблемы автоматической обработки естественного языка. Распознавание и синтез речи. Автоматический анализ и синтез речи. Методы синтеза. Устройство TTS-синтезатора речи. Автоматическое распознавание речи. Лингвистический и статистический подходы к распознаванию речи. Автоматическая обработка устной речи. Особенности диалога на естественном языке. Тест Тьюринга. Архитектура диалоговых систем. Чат-боты. Язык AIML. Вопросно-ответные системы.	
1.4	Системы автоматической обработки письменного текста.	Автоматический анализ естественного языка. NLP (Natural Language Processing). Автоматическая обработка письменного текста. Автоматическое распознавание текста. Автоматическая генерация текстов на ЕЯ.	

		Автоматический анализ и синтез текстов на естественном языке. Морфологический, синтаксический и семантический анализ в системах автоматической обработки текста. Синтаксический парсинг.	
1.5	Аспекты интеллектуального анализа текстов. Text Mining. Применение методов машинного обучения в лингвистике. Методы автоматической классификации	Текст как объект интеллектуального анализа. Классификация и кластеризация текстов. Алгоритмы классификации с учителем / без учителя. Векторизация. Конвейеры векторизации и преобразования. Машинное обучение в лингвистике. Формализация задач машинного обучения. Методы машинного обучения. Самообучающиеся системы: нейронные сети. Глубокое обучение. Способы автоматизации сбора информации из интернет-источников. Обучение моделей на размеченных текстах.	
1.6	Извлечение информации из текстов. Автоматическое аннотирование и реферирование текста.	Автоматизация подготовки и редактирования текстов на ЕЯ. Автоматы, формальные грамматики и языки. Контекстно-зависимый анализ текста. Извлечение информации из текстов. Распознавание сущностей. Автоматическое выявление метафор. Методы разрешения семантической неоднозначности. Разрешение анафоры и кореферентности. Установление референта. Использование экстратекстуальных сигналов. Извлечение отношений и событий. Зависимость категории от контекста.	
1.7	Теория и практика информационно-поисковых систем	Информационный поиск. Основные понятия информационного поиска. Индекс. Разработка информационно-поисковых систем. Классификация запросов. Релевантность, полнота, точность. Фильтрация и ранжирование. Факторы ранжирования. TF-IDF. Лингвистические аспекты информационного поиска. Лингвистическое обеспечение информационно-поисковых систем. Современные информационно-поисковые системы: принципы организации. Типология ИПС.	
1.8	Анализ тональности текста	Выделение мнений. Подходы к автоматическому определению тональности текста. Подход с использованием правил и словарей. Подход с использованием машинного обучения. Алгоритмы автоматического анализа тональности текста. Оценка качества работы алгоритмов. Сферы применения данных о тональности текста. Программы для определения тональности текста.	
1.9	Квантитативная лингвистика	Квантитативная лингвистика: основные понятия и методы описания языковых явлений. Структурно-вероятностная модель языка. Применение статистических данных в лингвистических	

		исследованиях. Статистическая стилистика. Квантитативная морфология. Закон Ципфа; коэффициент Жуйана. Частотные словари (в т.ч. общая частота/irm). Формальная семантика. Дистрибутивная семантика.	
1.1 0	Корпусная лингвистика. Лингвистические и филологические ресурсы и программы. Big Data	Корпусная лингвистика: история возникновения и развития. Принципы организации корпуса. Основные свойства корпуса. Классификации корпусов. Средства разметки. Типы разметки. Корпусные менеджеры. Современные корпуса текстов. Интернет как корпус. Мультимодальная лингвистика. Полевая лингвистика. Лингвистические ресурсы. Лингвистические базы данных. Построение и применение лингвистических ресурсов. Тезаурусы и онтологии. Обзор лингвистических баз данных. Примеры создания лингвистических ресурсов. Полуавтоматическое создание размеченного корпуса. Открытые данные и проблема их обработки. Верификация данных. Большие данные в филологии и языкознании.	
1.1 1	Компьютерная лексикография.	Компьютерная лексикография. Прикладные аспекты лексикографии. Электронные словари: основные типы словарей и принципы их организации. Структура словарной статьи электронного словаря. Электронные словари, доступные в сети. Частотные словари. Иноязычные словари. Лексическая информация в системах ИИ. Терминоведение и терминография: основные направления деятельности. Прикладные аспекты терминоведения. Теория письма; транскрипция; транслитерация.	
1.1 2	Автоматический перевод. Машинный перевод	Перевод как прикладная лингвистическая дисциплина. Виды перевода. Прикладные аспекты перевода. Лингвистические и нелингвистические проблемы перевода. Автоматический перевод. История создания систем автоматического перевода. Современное состояние отрасли. Машинный перевод. Машинный vs автоматизированный перевод. Определение машинного перевода. Основные подходы к машинному переводу. Основные способы перевода при помощи правил. Главная формула перевода. Модель перевода. Строение машинного перевода. Нейронный машинный перевод. Трансферный подход. Гибридный перевод. Методы оценки качества перевода. Современные системы машинного перевода. Актуальные проблемы машинного перевода. Перспективы развития машинного перевода. Методы оценки качества перевода. Сравнение систем машинного перевода.	
1.1	Лингвистическая	Лингвистическая экспертиза: виды и области	

3	экспертиза	применения. Экспертные системы. Экспертиза авторства текста. Стилеметрия. Дешифровка. Глоттохронология. Лингвокриминалистика. Психолингвистика. Когнитивная лингвистика. Нейролингвистика. Автоматизированные системы психолингвистического анализа текста. Фоносемантика. Компьютерная текстология. Лингвистическая прагматика. Классификации речевых актов. Вероятностные тематические модели. Социолингвистика. Политическая лингвистика. Речевое воздействие. Языковое манипулирование. Прикладные аспекты речевого воздействия. Семиотика. Основные понятия текстологии. Автоматическое сравнение рукописей. Компьютерная классификация рукописей.	
2. Лабораторные работы			
2.1	Системы автоматической обработки письменного текста.	Форматы данных и кодировки. Регулярные выражения. Макросы (Sublime, MS Excel). Распознавание и графематический анализ текста (OCR: ABBYY Fine Reader Online, New OCR, Online OCR). Сбор информации из интернет-источников. Data Mining. Автоматизация решения ДН-задач.	https://edu.vsu.ru/course/view.php?id=23349
2.2	Автоматическое аннотирование и реферирование текста	Визуализация данных (диаграммы, гистограммы, графики; облака слов, voyant tools, Flourish studio, Tableau). Графовые методы анализа текста, сетевой анализ (Gephi). Современные автоматизированные средства визуализации данных.	https://edu.vsu.ru/course/view.php?id=23349
2.3	Квантитативная лингвистика	Применение статистических методов в лингвистических исследованиях. Частотность. Закон Ципфа. Стоп-слова. Ipm.	https://edu.vsu.ru/course/view.php?id=23349
2.4	Корпусная лингвистика	Характеристика наиболее известных корпусов. Национальный корпус русского языка и другие русскоязычные корпусы. Корпусы иных языков. Обработка корпусных данных. Работа с параллельными корпусами (OPUS, Linguae, НКРЯ). Структура и назначение параллельных корпусов в решении задач ИИ. Google books Ngram Viewer. Выравнивание (Skuuper Cleaner, YouAlign, Champollion, Hunalign, LEOBILINGUA). Корпусные приложения (N-gram Viewer, SketchEngine), корпусные менеджеры (AntConc). Разметка. TEI. Именованные сущности. Markdown. BaseX. Язык запросов XPath	https://edu.vsu.ru/course/view.php?id=23349
2.5	Лингвистические ресурсы	Базы данных. Создание каталога лингвистических ресурсов	https://edu.vsu.ru/course/view.php?id=23349
2.6	Машинный перевод	Сравнение систем машинного перевода (Google Translate, Bing Translator, Yandex.Translate, Prompt)	https://edu.vsu.ru/course/view.php?id=23349

			se/view.php?id=23349
2.7	Компьютерная лингвистика как раздел прикладной лингвистики	Стемминг. Лемматизация. Topic Modelling. Текстовая близость. Классификация и кластеризация. Python	https://edu.vsu.ru/course/view.php?id=23349
2.8	Мультимодальная лингвистика. Полевая лингвистика	Работа с видео и аудио (Shotcut, ELAN, Praat)	https://edu.vsu.ru/course/view.php?id=23349
2.9	Лингвистическая экспертиза	Стилеметрия (в R)	https://edu.vsu.ru/course/view.php?id=23349
2.10	Системы автоматической обработки звучащей речи.	Обработка звучащей речи: программы анализа речи, программы синтеза речи. Чат-боты.	https://edu.vsu.ru/course/view.php?id=23349
2.11	Введение в NLP (Natural language processing)	Введение в GATE и извлечение информации. JAPE. GATE Cloud services. Машинное обучение в GATE. Python GateNLP. Opinion Mining. Анализ социальных сетей. Анализ оскорблений и дезинформации в онлайн-среде.	https://edu.vsu.ru/course/view.php?id=23349
	Защита проекта		

* заполняется, если отдельные разделы дисциплины изучаются с помощью онлайн-курса. В колонке Примечание необходимо указать название онлайн-курса или ЭУМК. В других случаях в ячейки ставятся прочерки.

13.2. Темы (разделы) дисциплины и виды занятий

№ п/п	Наименование темы (раздела) дисциплины	Виды занятий (количество часов)				
		Лекции	Практические	Лабораторные	Самостоятельная работа	Всего
1.1	Прикладная лингвистика как отрасль научного знания. Компьютерная лингвистика как раздел прикладной лингвистики	6		4	8	18
1.2	Системы автоматической обработки звучащей речи. Диалоговые системы и чат-боты	4		2	6	12
1.3	Системы автоматической обработки письменного текста.	4		2	6	12
1.4	Аспекты интеллектуального анализа текстов. Text Mining. Применение методов машинного обучения в лингвистике. Методы автоматической классификации. Введение в NLP	4		6	6	16
1.5	Извлечение информации из текстов. Автоматическое аннотирование и реферирование текста.	2		2	6	10
1.6	Теория и практика информационно-поисковых систем	2		2	6	10
1.7	Анализ тональности текста	2		2	6	10

1.8	Квантитативная лингвистика	2		2	6	10
1.9	Корпусная лингвистика. Лингвистические и филологические ресурсы и программы. Big Data	2		4	6	12
1.10	Компьютерная лексикография.	2		2	6	10
1.11	Автоматический перевод. Машинный перевод	2		2	6	10
1.12	Лингвистическая экспертиза. Мультимодальная лингвистика. Полевая лингвистика	2		4	8	14
	Итого:	34		34	76	144

14. Методические указания для обучающихся по освоению дисциплины

Необходимо регулярное посещение лекционных и лабораторных занятий, работа с литературой по дисциплине, выполнение индивидуальных лабораторных работ. Самостоятельная работа обучающихся предусматривает подготовку к аудиторным и лабораторным занятиям; выполнение домашних заданий; подготовку презентаций.

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

а) основная литература:

№ п/п	Источник
1	Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.
2	Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. — М.: Изд-во НИУ ВШЭ, 2017.
3	Леонтьева Н. Н. Автоматическое понимание текстов: системы, модели, ресурсы / Н. Н. Леонтьева. - М. : Академия, 2006. – 302 с.

б) дополнительная литература:

№ п/п	Источник
4	Баранов А. Н. Лингвистическая экспертиза текста. Теоретические основания и практика : учебное пособие / А. Н. - 2-е изд. - М. : Флинта : Наука, 2009. - 591 с.
5	Баранов А. Н. Лингвистическая экспертиза текста. Теоретические основания и практика : учебное пособие / А. Н. Баранов. - 5-е изд. - Москва : Флинта : Наука, 2013. - 591 с.
6	Гуслякова А. В. Информационные технологии и лингвистика XXI века : учебное пособие / А. В. Гуслякова. - Москва : МПГУ, 2016. - 96 с. - URL: http://biblioclub.ru/index.php?page=book&id=469675 (05.08.2019).
7	Баранов А. Н. Введение в прикладную лингвистику / А.Н. Баранов; Моск. гос. ун-т им. М.В. Ломоносова. Фил. фак. — М.: Эдиториал УРСС, 2001. — 358 с.
8	Шилихина, К.М. Основы прикладной лингвистики : учебное пособие по специальности 021800 (031301) - Теоретическая и прикладная лингвистика / К.М. Шилихина ; Воронеж. гос. ун-т. — Воронеж : ЛОП ВГУ, 2006. — 51 с.
9	Воеводская О. М. Информационные технологии и ресурсы Интернета в профессиональной деятельности переводчика : учебное пособие / О. М. Воеводская. - Воронеж : Издательский дом ВГУ, 2018 - URL: http://www.lib.vsu.ru/elib/texts/method/vsu/m18-92.pdf .
10	Всеволодова А.В. Компьютерная обработка лингвистических баз данных: учебное пособие / А.В. Всеволодова. — 2-е изд., испр. — М. : Флинта : Наука, 2007. — 90 с.
11	Герд А. С. Прикладная лингвистика / А.С. Герд ; С.-Петерб. гос. ун-т. — СПб. : Изд-во С.-Петерб. ун-та, 2005. — 266 с.
12	Жданов А. А. Автономный искусственный интеллект / А. А. Жданов. - 2-е изд. - М. : БИНОМ. Лабораторий знаний, 2009. - 359 с.
13	Захаров, В.П. Корпусная лингвистика. Учебник для студентов гуманитарных вузов / В.П. Захаров ; Богданова С. Ю. — Иркутск : Иркутский государственный лингвистический университет, 2011. — 161 с.

	// URL: http://biblioclub.ru/index.php?page=book&id=89753
14	Захарова Т. В. Практические основы компьютерных технологий в переводе : учебное пособие / Т. В. Захарова, Е. В. Турлова. - Оренбург : Оренбургский государственный университет, 2017. - 109 с. - URL: http://biblioclub.ru/index.php?page=book&id=481823 .
15	Зиятдинова Ю. Н. Теория перевода : курс лекций : учебное пособие / Ю. Н. Зиятдинова, Э. Э. Валеева. - Казань : Издательство КНИТУ, 2009. - 118 с. - <URL: http://biblioclub.ru/index.php?page=book&id=259076 >.
16	Зубов А. В. Информационные технологии в лингвистике / А.В. Зубов, И.И. Зубова. — М.: Academia, 2004. — 205 с.
17	Калмыков А.А., Коханова Л.А. Интернет-журналистика. – М.: ЮНИТИ-ДАНА, 2005. – 383 с.
18	Копотев М. Введение в корпусную лингвистику / М. Копотев. - Прага : Animedia Company, 2014. - 195 с. - URL: http://biblioclub.ru/index.php?page=book&id=375463 (05.08.2019).
19	Ляшевская О. Н. Корпусные инструменты в грамматических исследованиях русского языка. / О.Н. Ляшевская. - Москва : Издательский Дом ЯСК : Рукописные памятники Древней Руси, 2016. - 520 с.
20	Маннинг Кристофер Д. Введение в информационный поиск = Introduction to Information retrieval / Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце ; [пер. с англ. Д.А. Ключина] .— М. ; СПб. ; Киев : Вильямс, 2011. - 520 с.
21	Марчук Ю. Н. Компьютерная лингвистика : учебное пособие / Ю.Н. Марчук. - М. : Восток-Запад, 2007. - 317 с.
22	Моисеева И. Ю. Квантитативная лингвистика и новые информационные технологии : учебное пособие / И. Ю. Моисеева. - Оренбург : Оренбургский государственный университет, 2017. - 103 с. - URL: http://biblioclub.ru/index.php?page=book&id=481797 (07.08.2019).
23	Новое в зарубежной лингвистике. Вып. 24. Компьютерная лингвистика. – Москва: Прогресс, 1989. – 432 с. // URL: http://biblioclub.ru/index.php?page=book&id=38638
24	Новожилова, А.А. Информационные технологии в переводе. - Учебно-метод. пособие / , А.А. Новожилова, Е.В. Степанова, Е.А. Шовгенина - Волгоград: Изд-во ВолГУ, 2012. – 162 с.
25	Онтологии и тезаурусы: модели, инструменты, приложения : учебное пособие / Б. В. Добров, В. В. Иванов, Н. В. Лукашевич, В. Д. Соловьев. - Москва : Интернет-Университет Информационных Технологий, 2009. - 173 с. - URL: http://biblioclub.ru/index.php?page=book&id=233056 (07.08.2019).
26	Потапова Р. К. Новые информационные технологии и лингвистика: учебное пособие для студ. вузов / Р.К. Потапова; Моск. гос. лингв. ун-т. — Изд. 2-е. — М.: Едиториал УРСС, 2004. — 317 с.
27	Новожилова, А.А. Информационные технологии в переводе. - Учебно-метод. пособие / , А.А. Новожилова, Е.В. Степанова, Е.А. Шовгенина - Волгоград: Изд-во ВолГУ, 2012. – 162 с.
28	Потапова Р. К. Речь: коммуникация, информация, кибернетика: Учебное пособие для студентов вузов, обуч. по специальностям "Автоматизированные системы обработки информации и управления", "Лингвистика" / Р.К.Потапова. — М.: УРСС, 2001. — 564 с.
29	Прикладная и компьютерная лингвистика / Под ред. И.С. Николаева, О.В. Митрениной, Т.М. Ландо. – Москва: URSS, 2017. – 320 с.
30	Теория и практика машинного перевода : учебное пособие / авт.-сост. Э.В. Пиванова - Ставрополь : СКФУ, 2014. - 115 с. - URL: http://biblioclub.ru/index.php?page=book&id=457763
31	Титов В. Т. Романская квантитативная лексикология : (материалы к спецкурсу) : [учебно-методическое пособие для студентов 4 курса дневного и вечернего отделений] / В. Т. Титов ; Воронеж. гос. ун-т.- Воронеж : ЛОП ВГУ, 2006. - 43 с.
32	Berzins, K., Hudson, A. The Use of E-resources: A snapshot of e-resource use among Linking London LLN partner institutions. – London: University of East London, 2011. – URL: http://www.bbk.ac.uk/linkinglondon/resources/esystems-downloads/report_January2011_The_Use_of_Eresources_among_Linking_London_partners_Continuum.pdf
33	Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.
34	Blei D. (2012) Introduction to Probabilistic Topic Models // Communications of the ACM. — С. 77–84.
35	Burrows, J. (2002) 'Delta': a measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing, 17:267-87
36	Eder, M. Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. In: "Digital Humanities 2013: Conference Abstracts". University of Nebraska-Lincoln, Lincoln, NE, pp. 487-89.
37	Eder M. (2011) Style-markers in authorship attribution: A cross-language study of the authorial fingerprint. Studies in Polish Linguistics, 6:99–114.
38	Hoover, D. (2007) The End of the Irrelevant Text: Electronic Texts, Linguistics, and Literary Theory. Digital Humanities Quarterly 1.2.
39	Juola, P. (2006). Authorship Attribution. Foundations and Trends in Information Retrieval
40	Juola, P., Baayen, H. (2005) A Controlled-corpus Experiment in Authorship Identification by CrossEntropy, Literary and Linguistic Computing 20 (Suppl 1), pp. 59-67
41	Kestemont, M. (2014) Function words in authorship attribution. from black magic to theory? In Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL), pages 59–66, Gothenburg, Sweden.

42	Lieberman E., Michel, J.B., Jackson J., Tang T., Nowak M. (2007) Quantifying the Evolutionary Dynamics of Language. <i>Nature</i> . p. 449.
43	Michel, J.B., (2011). Quantitative analysis of culture using millions of digitized books . <i>Science</i> , 331(6014): pp. 176–82.
44	Mosteller F., Wallace D., (1963) Inference in an Authorship Problem. <i>Journal of the American Statistical Association</i> , Volume 58, Issue 302, pp. 275 - 309.
45	Newman M. (2010) <i>Networks: An Introduction</i> . Oxford: Oxford University Press.
46	Воронцов К.В. Вероятностное тематическое моделирование // http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf , 2013.
47	Мухин М. Ю. Лексическая статистика и идиостиль автора: корпусное идеографическое исследование : на материале произведений М. Булгакова, В. Набокова, А. Платонова и М. Шолохова: диссертация на соискание ученой степени доктора филологических наук. [Место защиты: ГОУВПО "Уральский государственный университет"]. - Екатеринбург, 2011. - 383 с.
48	Hamilton, William L., Jure Leskovec, and Dan Jurafsky. "Diachronic word embeddings reveal statistical laws of semantic change." <i>arXiv preprint arXiv:1605.09096</i> (2016).
49	Frermann, Lea, and Mirella Lapata. "A Bayesian Model of Diachronic Meaning Change." <i>TACL4</i> (2016): 31-45.
50	Stewart, Ian, Dustin Arendt, Eric Bell, and Svitlana Volkova. "Measuring, Predicting and Visualizing ShortTerm Change in Word Representation and Usage in VKontakte Social Network." <i>arXiv:1703.07012</i> (2017)
51	Hall, David, Daniel Jurafsky, and Christopher D. Manning. "Studying the history of ideas using topic models." In <i>Proceedings of the conference on empirical methods in natural language processing</i> , pp. 363-371. Association for Computational Linguistics, 2008.
52	Warner, Julian. <i>Human information retrieval</i> . Cambridge, MA: MIT Press, 2010.
53	Bryant, Fred B., and Paul R. Yarnold. "Principal-components analysis and exploratory and confirmatory factor analysis." (1995).
54	Wasserman, Stanley, and Katherine Faust. <i>Social network analysis: Methods and applications</i> . Vol. 8. Cambridge university press, 1994.
55	Lazer, David, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis et al. "Life in the network: the coming age of computational social science." <i>Science</i> (New York, NY) 323, no. 5915 (2009): 721.
56	Андреев В.С. Моделирование индивидуального стиля (на основе лингвистических характеристик): монография. М.: Флинта – Наука, 2012.
57	Введение в электронные лингвистические ресурсы [Электронный ресурс] / сост. В. Е. Гольдин, О. Ю. Крючкова. Саратов: 2011.
58	Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. <i>Онтологии и тезаурусы: модели, инструменты, приложения</i> . Изд-во ИНТУИТ, 2009.
59	Зубов А.В., Зубова И.И. <i>Основы искусственного интеллекта для лингвистов: учебное пособие для вузов</i> . Москва: Логос, 2007.
60	Леонтьева Н.Н. <i>Автоматическое понимание текстов: системы, модели, ресурсы</i> . М., 2006
61	Мартыненко Г.Я. <i>Основы стилеметрии</i> . Л.: Изд-во Ленинградского ун-та, 1988.
62	Марусенко М.А. <i>Атрибуция анонимных и псевдонимных литературных произведений методами теории распознавания образов</i> . Л.: Изд-во Ленинградского ун-та, 1990.
63	Игнатов Г., Михальча Р. <i>Текст Майнинг. Интеллектуальный анализ текста. Дизайн исследований, сбор данных и методы анализа / Гуманитарный центр</i> , 2021.
64	Осипов Г. С. <i>Методы искусственного интеллекта</i> . М.: Физматлит, 2011. Паттерсон, Д. <i>Глубокое обучение с точки зрения практика / Д. Паттерсон, А. Гибсон</i> . — Москва: ДМК Пресс, 2018.
65	<i>Прикладная и компьютерная лингвистика / Под ред. И.С. Николаева, О.В. Митрениной, Т.М. Ландо</i> . — Москва: URSS, 2017. — 320 с.
66	<i>Survey of Text Mining I: Clustering, Classification, and Retrieval / Ed. by M. W. Berry</i> . — 2004. — Springer, 2003. — 261 p.
67	Aggarwal C. C., Zhai C. <i>Mining Text Data</i> . — Springer, 2012. — 527 p.
68	Пескова О. В. Алгоритмы классификации полнотекстовых документов // <i>Автоматическая обработка текстов на естественном языке и компьютерная лингвистика</i> . — М.: МИЭМ (Московский государственный институт электроники и математики), 2011. — С. 170—212.
69	Kotu V., Deshpande B. <i>Text Mining // Data Science: Concepts and Practice (Second Edition)</i> , 2019. — P. 281-305.
70	Kushwaha A.K., Kar A.K., Ilavarasan P.V. Predicting retweet class using deep learning // <i>Trends in Deep Learning Methodologies: Algorithms, Applications, and Systems. Hybrid Computational Intelligence for Pattern Analysis</i> , 2021 – P. 89-112.
71	Nettleton D. <i>Text Analysis // Commercial Data Mining. Processing, Analysis and Modeling for Predictive Analytics Projects</i> , 2014, - P. 171-179

72	Silge J., Robinson D. Text Mining with R // O'Reilly Media, 2017. – 194 p.
73	Федюшкин Н.А., Федосин С.А. Краткий обзор методов и моделей интеллектуального анализа текста // Проблемы и достижения в науке и технике, 2017.
74	Chang, Angel X., Manolis Savva, and Christopher D. Manning. Semantic parsing for text to 3d scene generation. // ACL 2014: 17.
75	Усталов Д., Кудрявцев А. Применение онтологии при синтезе изображения по тексту. // Доклады всероссийской научно–практической конференции Анализ Изображений, Сетей и Текстов. М.: Национальный Открытый Университет ИНТУИТ. 2012
76	Wang H., Can D., Kazemzadeh A., Bar F., Narayanan S. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. // Proceedings of the ACL 2012 System Demonstrations (pp. 115-120). Association for Computational Linguistics.
77	Coppersmith G., Kelly E. Dynamic Wordclouds and Vennclouds for Exploratory Data Analysis // Sponsor: Idibon, 22, 2014.
78	Liu S., Wang X., Chen J., Zhu J., Guo B. TopicPanorama: A full picture of relevant topics. // Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on (pp. 183-192). IEEE.
79	Alexander E., Kohlmann J., Valenza R., Witmore M., Gleicher M. Serendip: Topic model-driven visual exploration of text corpora. // Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on (pp. 173-182). IEEE.
80	Smith A., Chuang J., Hu Y., Boyd-Graber J., Findlater L. Concurrent Visualization of Relationships between Words and Topics in Topic Models // Sponsor: Idibon, 2014, 79.
81	Заславская О.Ю., Пучкова Е.С. Подходы к визуализации объемных текстовых документов // Вестник Российского университета дружбы народов. Серия: Информатизация образования. 2016. №3.

в) информационные электронно-образовательные ресурсы (официальные ресурсы интернет)*:

№ п/п	Ресурс
82	Расширение Web Scraper для Chrome: http://webscraper.io/
83	https://voyant-tools.org/
84	Программное обеспечение в области автоматической обработки текста. www.aot.ru
85	Русский ворднет http://wordnet.ru
86	Технологии Академии «Яндекс» http://company.yandex.ru/technologies/search
87	Системы автоматического аннотирования текстов www.copernic.com
88	Тезаурус английского языка WordNet http://wordnet.princeton.edu
90	База данных языков мира SIL International. URL: http://www.sil.org/
91	Национальный корпус русского языка. URL: www.ruscorpora.ru
92	The Corpus of Contemporary American English. URL: http://corpus.byu.edu
93	Лингвистика Интернета: формирование дисциплинарной парадигмы (http://www.textology.ru/article.aspx?ald=76)
94	Новые возможности лингвистических исследований по исторической семантике с применением электронных ресурсов (http://textualheritage.org/content/view/74/68/lang.ru..)
95	Языковые ресурсы: традиции и инновации (http://elib.grsu.by/katalog/161659-346552.pdf)
96	Контроль использования интернет-ресурсов (http://alexott.net/ru/writings/cf/J1200502.pdf)
97	Компьютерная лингвистика (http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvi..)
98	ICT (http://iite.unesco.org/pics/publications/en/files/321..)
99	Подбор слов с помощью тезауруса. Режим доступа: https://support.office.com/ru-ru
100	Инструкция по использования программ конкорданс. Режим доступа: https://eflnotes.wordpress.com/2013/03/06/building-your-own-corpus-textstat-antconc/
101	British National Corpus. http://www.natcorp.ox.ac.uk/
102	Атлас языков мира. http://wals.info/
103	База данных «Языки мира» www.dblang.ru
104	Интернет-портал «Historia linguisticae». http://histling.nw.ru/
105	Коллекция словарей Института русского языка им. В.В. Виноградова РАН http://slovari.ru
106	Корпуса английского языка https://www.english-corpora.org/
107	Лингвистические данные Lingistic data consortium: https://www ldc.upenn.edu/language-resources
108	Лингвистический процессор «ЭТАП-3» http://proling.iitp.ru/ru/etap
109	Машинный фонд русского языка http://cfri.ru
110	Новый частотный словарь русской лексики. http://dict.ruslang.ru/freq.php
111	Программное обеспечение в области автоматической обработки текста. www.aot.ru
112	Проект «Вавилонская башня» http://starling.rinet.ru
113	Словари, созданные на основе Национального корпуса русского языка. http://dict.ruslang.ru/
114	Упсальский корпус русского языка. http://www.slaviska.uu.se/korpus.htm
115	Хельсинки аннотированный корпус русских текстов: slav.helsinki.fi/hanco

116	Чешский национальный корпус. http://ucnk.ff.cuni.cz/english/index.php
117	Каталог лингвистических ресурсов CLARIN https://www.clarin.eu/
118	Каталог лингвистических ресурсов ELRA http://www.elra.info/en/ .
119	Вестник Digital Humanities http://vdigital.me/
120	Программы лингвистического анализа и обработки текста http://asknet.ru/analytics/programms.htm
121	Программы анализа и лингвистической обработки текстов https://rvb.ru/soft/catalogue/c01.html
122	ACE - Automatic Content Extraction https://web.archive.org/web/20060308054306/http://www.itl.nist.gov/iad/894.01/tests/ace/
123	Интеллектуальный анализ текстов PROMT https://www.promt.ru/technology/text-analysis/
124	ТОП-5 инструментов для Text Mining http://datareview.info/article/top-5-instrumentov-dlya-text-mining/
125	Способы визуализации текстовой информации https://cs.hse.ru/vitext/visualize
126	Системы машинного перевода текстов и словари https://compress.ru/article.aspx?id=11757

* Вначале указываются ЭБС, с которыми имеются договора у ВГУ, затем открытые электронно-образовательные ресурсы, онлайн-курсы, ЭУМК

16. Перечень учебно-методического обеспечения для самостоятельной работы (учебно-методические рекомендации, пособия, задачки, методические указания по выполнению практических (контрольных), курсовых работ и др.)

№ п/п	Источник
127	Донина О.В. Автоматизация лингвистических исследований : учебное пособие для вузов / О.В. Донина. – Воронеж, 2022. – 125 с.
128	Донина О.В. Введение в анализ текстовых данных: учебное пособие для вузов / О.В. Донина, К.А. Сидоров, Н.С. Горбунов. – Воронеж, 2022. – 105 с.
129	Воеводская О. М. Информационные технологии в лингвистике [Электронный ресурс] : учебное пособие для вузов / О.М. Воеводская, И.А. Терентьева. - Воронеж : ИПЦ ВГУ, 2012 - <URL: http://www.lib.vsu.ru/elib/texts/method/vsu/m12-10.pdf >.
130	Донина О.В. Введение в большие языковые модели: учебник / О.В. Донина – Воронеж : Издательский дом ВГУ, 2024. – 132 с.
131	Донина О.В. Введение в фонетику: учебное пособие / О.В. Донина., К.А. Сидоров – Воронеж : Издательский дом ВГУ, 2022. – 126 с.
132	Коноплева, И.А. Информационные технологии : учебное пособие / И.А. Коноплева, О.А. Хохлова, А.В. Денисов ; под ред. И.А. Коноплевой. – 2-е изд., перераб. и доп. – Москва : Проспект, 2014. – 328 с. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=251652
133	Тимофеева, М.К. Язык с позиций философии, психологии, математики : учебное пособие : [16+] / М.К. Тимофеева. – 2-е изд., стер. – Москва : ФЛИНТА, 2013. – 176 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=375760
134	Волкова И.А. Введение в компьютерную лингвистику. Практические аспекты создания лингвистических процессоров. (Учебное пособие для студентов факультета ВМиК МГУ) // Издательский отдел факультета ВМиК МГУ (лицензия ЛР No040777 от 23.07.96), 2006 — 43 с.

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

При реализации дисциплины применяются технологии смешанного обучения (с использованием образовательного портала «Электронный университет ВГУ» (edu.vsu.ru)). Программа курса реализуется с применением дистанционных технологий.

Для выполнения практических заданий требуется следующее программное обеспечение: Microsoft Office, Praat, GitHub, Notepad++, Sublime, Hunalign, AntConc, Gephi, ELAN, BaseX, Java, R, R Studio, VoyantTools, GATE, KNIME, Orange, RapidMiner, wordsmith, nltk, Stanford coreNLP toolkit, word2vec.

18. Материально-техническое обеспечение дисциплины:

Для проведения лекционных и практических занятий используются аудитории, оснащенные специализированной мебелью.

Для самостоятельной работы используется класс с компьютерной техникой, оснащенный необходимым программным обеспечением, электронными учебными пособиями и законодательно – правовой и нормативной поисковой системой, имеющий выход в глобальную сеть.

19. Оценочные средства для проведения текущей и промежуточной аттестаций

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

№ п/п	Наименование раздела дисциплины (модуля)	Компетенция(и)	Индикатор(ы) достижения компетенции	Оценочные средства
1.	Прикладная лингвистика как отрасль научного знания. Компьютерная лингвистика как раздел прикладной лингвистики	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно-аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
2.	Системы автоматической обработки звучащей речи. Диалоговые системы и чат-боты	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно-аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
3	Системы автоматической обработки письменного текста.	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно-аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
4	Аспекты интеллектуального анализа текстов. Text Mining. Применение методов машинного обучения в лингвистике. Методы автоматической классификации. Введение в NLP	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно-аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
5	Извлечение информации из текстов. Автоматическое аннотирование и реферирование текста.	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно-аналитических	Устный опрос, выполнение индивидуальных лабораторных работ

№ п/п	Наименование раздела дисциплины (модуля)	Компетенция(и)	Индикатор(ы) достижения компетенции	Оценочные средства
			системах	
6	Теория и практика информационно- поисковых систем	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно- аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
7	Анализ тональности текста	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно- аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
8	Квантитативная лингвистика	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно- аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
9	Корпусная лингвистика. Лингвистические и филологические ресурсы и программы. Big Data	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно- аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
10	Компьютерная лексикография.	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно- аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
11	Автоматический перевод. Машинный перевод	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно- аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
12	Лингвистическая экспертиза. Мультимодальная лингвистика. Полевая лингвистика	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно- аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ

№ п/п	Наименование раздела дисциплины (модуля)	Компетенция(и)	Индикатор(ы) достижения компетенции	Оценочные средства
Промежуточная аттестация форма контроля – зачет с оценкой				<i>КИМ</i>

20 Типовые оценочные средства и методические материалы, определяющие процедуры оценивания

20.1 Текущий контроль успеваемости

Контроль успеваемости по дисциплине осуществляется с помощью следующих оценочных средств *Практикоориентированные задания/домашние задания, Сообщение/доклад/презентация, Устный опрос*

Примерный перечень тем рефератов

- Извлечение именованных сущностей и отношений.
- Дистрибутивная семантика.
- Тематическое моделирование.
- Стилеметрия.
- К проблеме анализа лингвистических данных при помощи компьютерных методов.
- Анализ текстов с использованием открытого программного обеспечения Voyant tools.
- Цифровые технологии в гуманитарном знании.
- Metadata в структуре объекта гуманитарных исследований. Opendata в гуманитарных исследованиях. Bigdata в гуманитарных исследованиях.
- Базы данных в гуманитарных исследованиях. Цифровые архивы.
- Системы машинного перевода: история разработок, нерешенные проблемы, перспективы. История создания систем машинного перевода
- Языковые корпуса: возможности практического применения в лингвистических исследованиях. Типы языковых корпусов
- Прикладные и теоретические лингвистические модели
- Применение систем автоматического распознавания речи
- Обработка речевого сигнала в системах распознавания речи.
- Применение систем автоматического анализа текста
- Моделирование семантики высказывания как прикладная лингвистическая проблема
- Онтологии. Семантические сети
- Автоматическое распознавание текста: основные принципы, проблемы и способы их разрешения
- Основы NLP.
- Text Mining.
- Требования, предъявляемые к системам представления и обработки знаний
- Семантические сети и графы. Фреймы
- Приобретение и формализация знаний. Трудности построения баз знаний
- Методы моделирования и обучения нейронных сетей
- Семантический анализ целого текста. Анализ тональности
- NLP в образовании и обучении языкам
- Автоматическая генерация учебных материалов
- Поддержка мультязычных учебных ресурсов
- ИИ-ассистенты для изучения языков
- ИИ-инструменты для изучения иностранных языков.

- Диалоговые боты как собеседники на иностранном языке
- Обратная связь на основе анализа ошибок пользователя
- Инструменты перевода и переформулировки
- Анализ прогресса обучающегося с помощью NLP.
- NLP и лингвистическое образование
- Автоматическое составление словарей частотности и коллокаций
- Тематическая сегментация текстов для учебных нужд
- Выделение морфологических и синтаксических паттернов
- Использование NLP в преподавании стилистики и риторики
- Обработка студенческих текстов и анализ ошибок
- Корректоры и помощники для написания текстов.
- Разработка лингвистических игр и тренажёров
- Геймификация.
- Использование NLP для построения игрового сюжета
- Создание чат-ботов-тренажёров
- Адаптивные языковые системы и обучение на данных
- Создание адаптивных языковых тренажёров на основе ИИ
- Создание ИИ-моделей, обучающихся на действиях учащегося
- Системы персонализированных рекомендаций
- Интеграция систем NLP в существующие платформы (Moodle, LMS)
- Оценка результатов и предложений от AI-помощника
- Этика и прозрачность в обучении с использованием AI
- NLP в цифровой гуманитаристике (Digital Humanities)
- Обработка художественной прозы
- Автоматический анализ литературных текстов: стиль, жанр, авторство.
- Автоматическое выделение фабулы и сюжета
- Тематическая сегментация текста
- Анализ взаимодействий персонажей (social graph)
- Выявление авторского стиля и жанровых особенностей
- Генерация краткого пересказа/аннотации
- Цифровая поэтика и поэзия
- Генерация поэтических текстов с заданными параметрами
- Интерпретация с помощью тематического моделирования
- Корпусная поэзия и визуализация
- Обработка исторических и архивных текстов
- Распознавание и нормализация старых орфографий
- Автоматическая разметка и анализ рукописей и архивных документов
- Аннотирование цифровых коллекций
- Визуализация исторических связей между документами
- Автоматическое выделение нарративных структур
- Генерация нарратива по заданной схеме
- Автоматический анализ парламентских и научных дебатов
- Обработка медиатекстов на предмет убеждающего воздействия
- Распознавание гендерно-маркированной лексики
- Сравнительный анализ социальных групп по языковым признакам
- Распознавание токсичности и агрессии
- Тематическое моделирование социальных конфликтов
- Инфраструктура и инструменты DH
- Интеграция NLP-инструментов в DH-платформы
- Разработка пользовательских интерфейсов для анализа текстов

- Цифровые библиотеки: построение и аннотирование коллекций.
- Использование графов, карт и визуализаций
- Автоматическая индексация архивов и рукописей
- Встроенные поисковые системы с NLP-поддержкой
- Системы совместной интерпретации и комментирования
- Стандарты представления данных (TEI, XML, RDF)
- Влияние NLP на современные гуманитарные исследования.
- Интеграция методов NLP в исследовательские платформы.
- Автоматическая аннотация, резюмирование и визуализация текста
- Обобщение и переформулирование текста
- Sentence compression: алгоритмы сжатия предложений
- Автоматическое переформулирование и перефразирование
- Генерация кратких версий с сохранением смысла
- Использование трансформеров для перефразирования
- Автоматическое реферирование и резюмирование
- Экстрактивные и абстрактивные методы
- Модели для генерации резюме (BART, PEGASUS и др.)
- Автоматическая генерация заголовков
- Сжатие многостраничных документов
- Качество резюме: метрики ROUGE, METEOR, BLEU
- Визуализация смысловой структуры текста
- Semantic graphs и topic maps
- Облачные платформы: Tableau, Voyant Tools, Flourish и др. облачные сервисы для визуализации смысловых связей.
- Визуализация ключевых тем, связей и сюжетов
- Интерактивные карты текста и нарратива
- Системы представления аргументации и риторики
- Сжатие и компактное представление больших текстов
- Обработка больших текстов: методы компактного представления.
- Алгоритмы уплотнения и фильтрации информации
- Сжатие диалогов, форумов, соцсетей
- Представление документов в векторных и графовых формах
- Методы повышения читаемости длинных текстов
- Автоматическая коррекция грамматических ошибок
- Сравнение правил и ML-подходов при автоматической коррекции ошибок
- Rule-based grammar checkers vs нейросетевые модели
- Ограничения ML в сложных грамматических конструкциях
- Коррекция пунктуации и стилистики
- Алгоритмы восстановления пунктуации (dot/zero punctuation recovery)
- Стилизация текста под жанр/аудиторию
- Обнаружение тавтологий и речевых штампов
- Многоязычные и малоресурсные языки
- Обучение моделей на синтетических данных
- Перенос моделей с английского на другие языки
- Построение корпусов ошибок для языков с низкими ресурсами
- Многоязычные корректоры и интерфейсы
- Анализ данных в гуманитарных исследованиях
- Подходы к анализу данных. Дизайн исследования
- Сравнение data mining, text mining и data science
- Критерии качества данных в гуманитарных науках.

- Построение корпуса: от сбора до чистки
- Аннотирование и категоризация данных
- Проблема интерпретации и репрезентативности в text mining.
- Методы анализа текстов. Примеры и кейсы
- Текстовая классификация и тематическое моделирование
- Кластеризация мнений, сюжетов, персонажей
- Обучение с учителем и без: возможности и риски
- Методы извлечения знаний и построения онтологий
- Визуализация результатов исследований
- Автоматический контент-анализ СМИ и документов
- Применение нейросетей в лингвистических задачах.
- Анализ рецензий и отзывов
- Сентимент-анализ
- Определение тональности текста (positive, neutral, negative)
- Категориальный анализ: цена, качество, доставка, сервис
- Выделение субъективных выражений и полярности
- Сравнение по языкам и жанрам (отзывы, посты, комментарии)
- Обнаружение фейковых отзывов
- Поведенческие и лексические признаки фейков
- Обучение моделей на размеченных фейковых данных
- Сравнение моделей: ML vs rule-based
- Выделение аспектов и ключевых тем
- Определение тематики отзыва
- Модели аспектно-ориентированного анализа (ABSA)
- Сравнение моделей оценки отзывов
- Извлечение упоминаемых товаров, услуг, характеристик
- Картирование лексических полей вокруг аспектов
- Агрегация мнений и генерация обзоров
- Сбор отзывов и построение метаобзора
- Генерация обобщающего текста
- Статистическое представление мнений (графики, рейтинги)
- Автоматическое обновление агрегированных обзоров
- Анализ текстовых данных с помощью Python и библиотек NLP
- Работа с Gensim для анализа текстов
- Введение в Gensim и его возможности для обработки текста: назначение и основные модули
- Построение корпуса и словаря (Dictionary, Corpus) с использованием Gensim
- TF-IDF: формирование и интерпретация матрицы для оценки важности слов
- Построение моделей Word2Vec, FastText, Doc2Vec
- Оценка семантической близости и сходства между текстами с помощью метрик Gensim.
- LDA-моделирование тем (topic modeling)
- Визуализация тематической структуры текста
- Обработка текста с NLTK и spaCy
- Токенизация, лемматизация, стемминг
- Определение частей речи и синтаксический разбор
- Именованные сущности (NER) и их классификация
- Анализ синтаксической зависимости (Dependency Parsing)
- Создание простых пайплайнов обработки текста
- Сравнение возможностей NLTK и spaCy

- Масштабируемый анализ с PySpark
- Введение в Apache Spark и NLP на больших данных
- Работа с RDD и DataFrame для текстовых коллекций
- Распределённая токенизация и обработка
- Использование распределённых вычислений для обработки текстов.
- Реализация моделей Word2Vec и TF-IDF на PySpark
- Применение MLlib для классификации текстов
- Интеграция с Hadoop и облачными хранилищами
- Визуализация результатов анализа с помощью Python
- Построение графиков и диаграмм: matplotlib, seaborn, Plotly
- Визуализация частотности слов, биграмм и тем
- Текстовые облака и тепловые карты
- Визуализация структуры документа и токенов
- Интерактивные панели для представления результатов
- Будущее NLP: тренды и технологические горизонты
- Развитие мультязычных моделей
- Модели, работающие с десятками языков (mBERT, XLM-R)
- Проблемы и вызовы перевода и кросс-языкового переноса
- Роль low-resource языков в развитии NLP
- Многоязычные датасеты и корпуса
- Персонализация и демократизация NLP
- Создание пользовательских чат-ботов без кода
- Персонализированные голосовые ассистенты
- Настройка систем под конкретного пользователя
- Эволюция диалоговых интерфейсов с помощью LLM
- Обучение моделей на локальных данных
- Интеграция NLP с другими технологиями
- NLP + AR/VR: описание и анализ пространственного текста
- Объединение NLP с компьютерным зрением (multimodal models)
- Текст как интерфейс в виртуальных мирах
- Этические и социальные вызовы
- Проблема генерации фейков и дезинформации
- Снижение прозрачности принятия решений
- Этические нормы в генеративных системах
- Вопросы приватности и авторства
- Введение в большие языковые модели (LLM) и архитектуры трансформеров
- Архитектура трансформеров
- Encoder, decoder, self-attention
- Маскирование и позиционное кодирование
- Принцип многоуровневой обработки текста
- Механизмы памяти и масштабирования
- Отличие от RNN и CNN-моделей
- Обзор и сравнение LLM
- GPT (1–4), BERT, RoBERTa, T5, Claude, Mistral: Архитектурные различия и сферы применения. Сравнение по метрикам и задачам
- Преимущества автокодировщиков и автогенераторов
- Обучение и адаптация моделей
- Принцип обучения языковых моделей на больших корпусах.
- Fine-tuning для специфических задач (NER, QA, кластеризация)
- Transfer learning и zero-shot методы

- Использование специализированных датасетов
- Ограничения и вызовы масштабируемых языковых моделей.
- Инженерия запросов и дообучение
- Prompt engineering: шаблоны и стратегии
- Few-shot, zero-shot и in-context learning
- Роль температуры и длины вывода
- Do-it-yourself LLM: адаптация open-source моделей
- Ограничения и уязвимости при дообучении
- Применение LLM в разных областях. Примеры применения LLM в гуманитарных, социальных науках и прикладной лингвистике.
- Визуализация данных в цифровой лингвистике
- Языковые карты. Тематическая и семантическая визуализация
- Визуализация диалектных зон и языковой изменчивости
- Визуализация семантических сетей и соотношений
- Интерактивные карты и линейные диаграммы изменений
- Отображение тем и ключевых слов
- Визуализация изменений тем и смыслов.
- Распределение значений во времени и жанрах
- Визуализация тем с LDA и BERTopic
- Семантические векторы и редукция размерности (t-SNE, PCA)
- Интерфейсы и дашборды
- Создание дашбордов с Plotly Dash, Streamlit
- Интерактивные панели и фильтрация данных
- Интерактивные интерфейсы анализа текстов.
- Обратная связь в виде визуализации (user-in-the-loop)
- Дашборды для анализа диалога и корпуса.
- Визуализация результатов анализа моделей
- Графики важности признаков
- Отображение ошибок классификации и вкладов токенов
- Визуальный анализ attention weights
- Диагностика и настройка моделей на основе визуальных признаков
- Визуализация текстовых данных и результатов анализа.
- Визуализация признаков и особенностей текста.
- Использование графиков для представления результатов анализа.
- Разработка интерактивных визуализаций для текста.
- Применение визуальных инструментов для анализа больших объемов данных.
- Сравнение типов диаграмм и их уместность
- Практики UX-дизайна для текстовой аналитики
- Гендер и язык в NLP: от анализа до этики. Гендерные стереотипы и предвзятость. Гендерная нейтральность и инклюзивность
- Методы выявления гендерных стереотипов в языковых моделях
- Гендерный bias в word embeddings и трансформерах
- Кейсы гендерной дискриминации в машинных генерациях
- Подходы к устранению и нейтрализации bias
- Генерация текста в гендерно-нейтральной форме. Поддержка инклюзивной лексики
- Алгоритмы редактирования гендерных маркеров. Платформенные решения (например, Microsoft Inclusive Writing)
- Разметка и анализ гендерных маркеров. Гендер в медиатекстах и социальных сетях. Сбор и разметка корпусов с гендерной информацией
- Определение гендерно маркированных форм в тексте
- Автоматическое распознавание гендера по лексике и контексту

- Корпусные и визуализационные подходы
- Генерация и анализ текста с помощью ИИ.
- Алгоритмы генерации текста
- Greedy search, beam search, top-k и nucleus sampling
- Температурные параметры генерации
- Комбинирование стратегий генерации
- Влияние алгоритмов на креативность и связность
- Модели генерации и их обучение
- Архитектуры генерации: GPT, T5, LLaMA
- Fine-tuning генеративных моделей
- RLHF и instruction tuning
- Переформулирование, резюмирование, сочинения и классификация текстов
- Метрики оценки и интерпретация
- BLEU, ROUGE, METEOR, perplexity: плюсы и минусы
- Когерентность, связность, правдоподобие
- Интерпретация «галлюцинаций» и несостыковок
- Визуализация внимания и логики вывода
- Управление длиной, стилем, темой генерации. Имитация авторского стиля
- Инструменты контролируемой генерации (CTRL, Z-Code): управление стилем и тоном
- Генерация: Этические риски (фейки, пропаганда, манипуляции). Борьба с bias и галлюцинациями
- Автоматизация отчётов, писем и маркетинга
- Глубокое обучение в NLP: архитектуры и применение
- Базовые методы машинного и глубокого обучения в NLP
- Классификация, регрессия, кластеризация
- Метод главных компонент (PCA) и t-SNE для визуализации
- RNN, LSTM, CNN: принципы работы и сравнение
- BERT и GPT как примеры трансформеров
- Архитектуры трансформеров
- Трансформеры: принцип работы и эволюция
- Self-attention и позиционное кодирование
- Encoder vs decoder и полные encoder-decoder системы
- Модели: BERT, RoBERTa, GPT, T5, Electra
- Применение трансформеров в задачах NER, QA, MT
- Fine-tuning и адаптация моделей
- Fine-tuning предобученных моделей под конкретные языковые задачи
- Fine-tuning на малых датасетах
- Инструкция и контекстуальное обучение (instruction/few-shot)
- Instruction Tuning.
- Сравнение zero-shot и few-shot обучения.
- LoRA, P-tuning, Adapter layers
- RLHF (обучение с подкреплением от обратной связи человека)
- Интерпретируемость нейросетевых моделей.
- Attention maps и их интерпретация
- Explainable AI (XAI) в NLP
- Анализ ошибок и вкладов токенов
- Метрики доверия и уверенности модели
- Метрики оценки качества моделей: F1, BLEU, METEOR, ROUGE, perplexity.
- Мультиязычность и сжатие

- Мультимодальные модели: CLIP, Flamingo, GPT-4V
- Дистилляция моделей и квантование
- Оптимизация моделей для мобильных и edge-устройств
- NLP-пайплайны и инструменты
- Нейросетевые архитектуры для лингвистических задач.
- Интеграция лингвистических признаков в нейросети
- Spark NLP: масштабируемая обработка текста
- Hugging Face Transformers: загрузка, настройка, инференс
- FastAPI и REST-сервисы для NLP
- Интеграция моделей в приложения и интерфейсы
- Автоматизация NLP пайплайнов
- Графовые методы для анализа текста
- Построение графов из текстовых данных
- Преобразование текстов в графовые структуры
- Графы слов, понятий, фраз, предложений и текста
- Этапы токенизации и лемматизации для графового анализа
- Использование co-occurrence и dependency parsing
- Сравнение синтаксических и семантических графов
- Методы анализа графов в NLP
- Метрики центральности (degree, betweenness, closeness)
- Поиск кластеров и сообществ в текстовых графах
- Анализ связности и плотности графа
- Методы вычисления и анализа графов для текстов.
- Идентификация ключевых слов и концептов
- Извлечение и визуализация графов
- Извлечение графов из текстов с использованием NLP
- Визуализация и интерпретация графовых структур.
- Построение графов с NetworkX и визуализация через Gephi
- Работа с форматами GraphML и GEXF
- Использование Neo4j для хранения и запроса текстовых графов
- Визуализация дискурсивных и семантических связей
- Прикладные задачи с графами
- Разрешение кореференции и сущностей через графы
- Анализ диалога и дискурса на графовом уровне
- Построение графов для систем рекомендаций
- Использование графов в задачах кластеризации текстов
- Графовые подходы в задачах факт-чекинга и QA
- Применение графов к задачам тематического моделирования
- Анализ дискурса через графы.
- Диалоговые системы и чат-боты
- Основы и архитектура чат-ботов. Сравнение rule-based и ML-ботов
- Архитектура диалоговых систем: intent, NLU, dialog manager
- Архитектура голосовых и текстовых агентов
- Обработка интенгов и слотов
- Диалоговые деревья и FSM
- Языки разметки и сценарии взаимодействия (AIML).
- Построение и обучение чат-ботов
- Обработка вопросов и генерация ответов
- Принципы построения диалоговых систем.
- Разработка сценариев: от скриптов до генерации

- Подготовка и «очистка» корпуса для обучения: нормализация, унификация формулировок, синтаксический фильтр.
- Практика проектирования чат-ботов на платформах.
- Обучение на основе примеров: RASA, Dialogflow, JAICP
- Дообучение с применением LLM для улучшения диалога
- Обучение ботов для рекомендаций и поиска информации
- Диалоговое поведение и контекст
- Основы проектирования диалогового взаимодействия.
- Распознавание речевых актов
- Обработка многопоточной беседы и долгосрочного контекста
- Управление историей диалога. Вежливость, эмпатия и социальные нормы
- Применение правил для построения вежливых и адекватных бесед.
- Имитация стилистических и ролевых особенностей
- Разработка чат-ботов для различных типов взаимодействий.
- Оценка качества работы чат-бота и улучшение его функционала.
- Этические аспекты и социальное поведение ботов.
- UX и оценка удовлетворённости пользователя
- Сценарии «честности» и «ухода от ответа». Антифрод и антиабыюзные механизмы
- Тест Тьюринга и вопросы имитации интеллекта
- Инструменты и мультимодальность
- Обзор платформ: Dialogflow, RASA, Microsoft Bot Framework
- Интеграция речевых технологий: ASR и TTS
- Разработка мультимодальных диалогов
- Интеграция мультимодальности: речевые, текстовые и визуальные реакции в диалогах
- Извлечение информации и анализ тональности
- Распознавание именованных сущностей (NER)
- Выделение сущностей и отношений.
- Извлечение ключевых слов.
- Выявление событий в тексте.
- Основные подходы: правила, CRF, BiLSTM, BERT
- Обработка NER в мультязычных корпусах
- Обогащение и нормализация сущностей (entity linking)
- Использование готовых моделей: spaCy, Flair, Stanza
- Визуализация извлечённых сущностей и связей
- Извлечение связей и событий
- Определение семантических отношений между сущностями
- Relation Extraction: подходы и модели
- Event Extraction: выявление событий и участников
- Работа с неявными отношениями и фреймами
- Кореференция и разрешение анафор
- Построение семантических графов связей
- Анализ тональности и субъективности
- Методы sentiment analysis: lexicon-based, ML и DL
- Aspect-based Sentiment Analysis (ABSA)
- Определение эмоциональной окраски текста
- Выделение мнений
- Методы opinion mining
- Выделение субъективных высказываний
- Распознавание иронии и сарказма

- Автоматическое выявление метафор и переносов.
- Разрешение анафоров и установление кореференции.
- Контекстно-зависимый анализ и семантическая неоднозначность.
- Инструменты и приложения анализа тональности
- Библиотеки и фреймворки: VADER, TextBlob, transformers
- Обработка отзывов, комментариев, социальных сетей. Мониторинг общественного мнения и брендов
- Анализ тональности в новостных и политических текстах. Применение в маркетинге, HR, PR, СМИ, соцсетях и психолингвистике
- Информационно-поисковые системы и текстовый поиск
- Основы построения поисковых систем
- Индексация текста: токенизация, нормализация, инвертированные индексы
- Построение поискового пространства: фильтрация и сжатие
- Введение в TF-IDF, BM25 и векторные модели
- Поиск по ключевым словам и фразам
- Методы предобработки запроса
- Языковая поддержка поисковых алгоритмов.
- Алгоритмы ранжирования и релевантности
- Метрики релевантности: precision, recall, MAP, nDCG
- Кластеризация и фильтрация результатов поиска
- Модели на основе обучения ранжированию (Learning to Rank)
- Интеграция пользовательских поведенческих факторов
- Нейросетевые модели ранжирования: ColBERT, DPR
- Семантический и мультимодальный поиск
- Автоматический вопросно-ответный поиск
- Семантические модели: DSSM, SBERT, semantic hashing
- Векторный поиск и embedding-based retrieval
- Интеграция изображений, графов и текста в поиске
- Multimodal QA и DocVQA
- Поддержка мультязычности и кросс-языкового поиска
- Вопросно-ответные и генеративные поисковые системы
- Open-domain QA: DPR, Retrieval-Augmented Generation (RAG)
- Extractive vs generative QA
- Построение пайплайнов: retriever + reader
- Примеры систем: Haystack, Elasticsearch + Transformers
- Визуальный поиск и ODQA-интерфейсы
- Инфраструктура для аналитических проектов: от Jupyter до Power BI
- Организация среды аналитика. Настройка Jupyter Notebook, Zeppelin, DataSpell
- Использование Conda, Docker, venv для изоляции окружений
- Организация структуры проектов и reproducibility. Подключение к БД и API из рабочих сред. Интеграция с Git, GitHub, DVC для контроля версий
- Расширения Jupyter и полезные плагины
- Управление пайплайнами и автоматизация
- Создание ETL и аналитических пайплайнов
- Обзор инструментов: Apache Airflow, Prefect, Dagster
- Параллельные и распределённые вычисления в пайплайне
- BI и визуализация результатов
- Сравнение BI-инструментов: Tableau, Power BI, Metabase, Superset
- Создание дашбордов и отчётов
- Визуализация времени, категорий, иерархий
- Взаимодействие BI с Python-скриптами и Jupyter

- Поддержка автоматических обновлений отчётов
- Использование R Studio для текстового анализа
- Настройка RStudio и подключение библиотек для обработки текстовых данных. Загрузка текстовых данных (csv, JSON, scraping). Работа с текстовыми векторами и списками
- Импорт и парсинг данных из сайтов (RCurl, rvest)
- Работа с кодировками и мультязычными данными
- Предобработка текста и text mining
- Очистка и нормализация текста (tm, tidytext)
- Удаление стоп-слов, лемматизация, стемминг
- Построение Document-Term Matrix
- Частотный анализ, n-граммы
- Визуализация частот (wordcloud, barplot)
- Сравнение подходов в R Studio и Python для обработки текстов.
- Визуализация данных и результатов анализа в R.
- Использование ggplot2 для графиков
- Визуализация кластеров, PCA, dendrogram
- Построение интерактивных графиков (plotly, shiny)
- Автоматизированная генерация отчётов (R Markdown)
- Построение диаграмм связи и распределений
- Квантитативная лингвистика и числовые модели языка
- Статистика в лингвистике
- Квантитативный анализ текстов: статистические методы обработки.
- Закон Ципфа, закон Менцерата, коэффициент Жуйана
- Частотные словари и кривые распределения
- Применение статистических методов для анализа частоты слов.
- Формализация лексической структуры языка.
- Статистические методы в корпусной лингвистике
- Меры разнообразия и плотности словаря
- Регрессия и корреляционный анализ в лингвистике
- Моделирование языка и стилометрия
- Статистическая стилистика: авторство, жанр
- Дистрибутивная семантика и контекстные векторы
- Квантитативные подходы к морфологии
- Анализ длины слов и предложений
- Расчёт индексов сложности текста (ARI, Flesch-Kincaid)
- Обработка больших текстовых данных
- Построение и анализ частотных таблиц
- Анализ распределения слов и фраз в текстах.
- Распределение частот в различных жанрах и стилях
- Визуализация закономерностей в текстах
- Применение кластеризации и PCA
- Классификация текста: подходы и инструменты
- Базовые алгоритмы классификации текста
- Разработка модели классификации текста. Создание приложений для автоматической классификации текстов.
- Классификация текстов с использованием статистических моделей.
- Наивный байесовский классификатор
- Логистическая регрессия и SVM
- Метод k ближайших соседей (k-NN)
- Сравнение простых моделей классификации

- Подходы к векторизации текста (BoW, TF-IDF)
- Построение baseline-моделей для классификации
- Классификация с применением глубокого обучения
- Использование нейросетей для классификации текстов
- Применение LSTM, BiLSTM, GRU
- Fine-tuning моделей BERT, RoBERTa
- Многоклассовая и бинарная классификация
- Аугментация и балансировка классов
- Обработка несбалансированных выборок
- Метрики оценки: Accuracy, Precision, Recall, F1
- Кросс-валидация и подбор гиперпараметров
- Интерпретация моделей (LIME, SHAP)
- Классификация по стилю, тематике, авторству
- Визуализация результатов классификации
- Основы кластеризации в NLP
- Различие между обучением с учителем и без
- Подготовка текстовых данных для кластеризации
- Векторизация текста для кластеризации
- Снижение размерности (PCA, t-SNE, UMAP)
- Проблема выбора числа кластеров
- Алгоритмы кластеризации
- K-means и его модификации
- DBSCAN и иерархическая кластеризация
- Кластеризация тематических моделей (LDA)
- Affinity Propagation, MeanShift
- Оценка результатов кластеризации
- Выявление семантических групп и тем
- Поиск аномалий в кластерах
- Визуализация кластеров (pyLDAvis, plotly, seaborn)
- Интерпретация кластеров и ключевых слов
- Применение кластеризации для выявления сходства в текстах.
- Когнитивные аспекты языковой обработки
- Моделирование когнитивных категорий
- Языковые категории и их реализация в NLP
- Прототипическая структура категории
- Категориальные модели в распределённой семантике
- Психолингвистические данные в NLP
- Язык и образ мира: моделирование на уровне лексики
- Автоматическое выявление метафор
- Метафора и перенос в нейросетевых моделях
- Метафорические модели и их распознавание
- Связь когнитивной лингвистики и нейросетей
- Соотнесение нейросетевых моделей с когнитивными схемами
- Концептуальная структура в embedding-моделях
- Интерпретируемость как когнитивная задача
- Нейросимволические подходы
- NLP для анализа образов мышления (прецеденты, сценарии)
- Автоматическая лексикография
- Извлечение и структурирование лексики
- Автоматическое извлечение лексем и значений

- Анализ частоты и сочетаемости слов
- Построение онтолексиконов и онтологий
- Технологии транслитерации и формального описания лексики.
- Лексикографические ресурсы и инструменты
- Применение словарей в задачах ИИ и NLP.
- Типология электронных словарей (частотные, толковые, билингвальные, терминологические) и их внутренняя структура
- Использование WordNet, FrameNet, Wiktionary, Open Multilingual Wordnet
- Форматы хранения и разметки: TEI, LMF, OntoLex
- Визуализация лексических связей (graph-based модели)
- Разработка и использование лексикографических ресурсов.
- Компьютерный анализ текста и создание лингвистических ресурсов
- Технологии и приложения в лексикографии
- Машинное обучение и deep learning в лексикографии
- Методы синтаксической и семантической фильтрации
- Использование annotated corpora (UD, Tatoeba)
- Автоматизация лексикографических процессов
- Методы работы с синтаксическими деревьями и аннотированными корпусами
- NLP-конвейеры и векторизация текста
- Архитектура NLP-конвейеров
- Концепция pipeline в NLP: последовательная обработка
- Конвейеры в Scikit-learn: Pipeline и FeatureUnion
- Разработка конвейеров для предобработки и анализа данных.
- Подключение внешних трансформеров и кастомных функций
- Использование spaCy и HuggingFace Pipelines
- Параллельная и распределённая обработка
- Методы векторизации текста
- Преобразование слов в векторные представления.
- Векторизация: подходы, библиотеки, применение.
- Bag of Words и TF-IDF: принципы, реализация
- Word2Vec и FastText: обучение и применение
- GloVe и Doc2Vec: особенности моделей
- BERT и контекстуальные эмбединги
- Оптимизация и масштабирование векторизации
- Векторизация больших коллекций текста
- Использование HashingVectorizer и сокращение размерности
- Кэширование результатов и конвейеризация
- Работа с Dask, Joblib, Spark NLP
- Эффективное хранение и сериализация векторных представлений
- Контекстно-зависимый анализ и n-граммы
- Работа с n-граммами
- Генерация n-грамм из текстов (2-граммы, 3-граммы и выше)
- Сравнение uni-, bi- и tri-грамм
- Частотный анализ и фильтрация n-грамм
- Использование n-грамм в классификации и кластеризации
- Создание языковых моделей на n-граммах
- Использование n-грамм для анализа контекста.
- Контекстуальный анализ на основе грамматических признаков
- Применение контекстуальных данных для улучшения анализа.
- Выделение POS-тегов и морфологических признаков

- Интеграция синтаксической информации в модели анализа
- Построение моделей для извлечения значений из текста.
- Использование n-граммных моделей для генерации фраз
- Построение вероятностных языковых моделей
- Извлечение информации через шаблоны и контексты
- Статистические и ML-подходы к извлечению значений
- Учет локального контекста в извлечении смыслов
- Применение грамматических признаков для извлечения информации.
- Корпусная лингвистика и обработка больших текстов
- Создание и аннотирование корпусов
- Сбор текстов и создание корпуса
- Методы аннотирования: морфология, синтаксис, семантика
- Использование форматов TEI, CoNLL, JSONL
- Применение разметки UD (Universal Dependencies)
- Инструменты аннотации: brat, WebAnno, INCEpTION
- Автоматическая лемматизация в корпусах.
- Морфосинтаксическая разметка на больших данных.
- Сравнительный анализ стилей и тематики в больших корпусах.
- Инструменты анализа корпусов
- Поисковые системы для корпусов.
- Применение AntConc, Sketch Engine, Voyant
- Использование NLTK, spaCy и других средств анализа
- Визуализация закономерностей в корпусе
- Инструменты Baleen, Sqlite и NLTK CorpusReader.
- Промежуточный анализ корпуса: частотные списки, describe(), визуализация.
- Частотные списки, KWIC и коллокации
- Поиск ключевых слов и шаблонов
- Работа с параллельными и специализированными корпусами
- Параллельные корпуса и их использование в NLP.
- Сбор и построение корпусов под задачи.
- Структура параллельных корпусов и alignment
- Создание терминологических и тематических подкорпусов
- Корпуса для обучения и тестирования NLP-моделей
- Веб-скрапинг и сбор текстов
- Основы Web Crawling: структура HTML, robots.txt.
- Использование BeautifulSoup, Scrapy, Selenium
- Парсинг HTML, XML, JSON
- Работа с API, RSS, выгрузка текстов из сайтов
- Этические аспекты автоматического сбора данных
- Очистка и нормализация неструктурированного текста
- Хранение, фильтрация и визуализация данных
- Хранение корпусов: базы данных, файловые форматы
- Использование Lucene, Elasticsearch для индексации
- Методы фильтрации контента
- Создание дашбордов и графиков (Plotly, Seaborn)
- Визуализация структурных и частотных характеристик текста
- Корпусная статистика и её визуализация.
- Краудсорсинг аннотаций: Яндекс.Толока, Prolific
- Проверка качества данных: перекрёстная валидация
- Стандарты разметки и их соблюдение

- Обзор популярных лингвистических корпусов в Интернете.
- Инструменты для работы с корпусами: текстовые анализаторы и парсеры.
- Инструменты для работы с большими корпусами текста.
- Трансформация данных корпуса для использования в моделях.
- Критический подход к «сырым» данным и интерпретируемости в NLP
- Понятие «сырых данных» в текстовом анализе
- Что такое «сырые» данные: типы и примеры
- Трудности работы с неразмеченными корпусами. Влияние форматов, кодировок и предобработки
- Языковые шумы: орфографические ошибки, неоднозначности
- Границы между «сырыми» и «обработанными» данными
- Скептицизм в эпоху Big Data
- Различие между шумом и значимыми паттернами
- Переобучение на редких феноменах в лингвистических данных
- Избыточное доверие статистике при малых объёмах
- Этические аспекты гиперинтерпретации результатов
- Когнитивные ловушки и ложные инсайты
- Эвристики, искажения и когнитивные ловушки в анализе.
- Ошибки подтверждения гипотез
- Случайные корреляции и их интерпретация
- Примеры ложных выводов в проектах NLP и как их избежать
- Кросс-корпусный и межъязыковой анализ
- Использование Universal Dependencies
- Интероперабельность аннотированных корпусов
- Семантическое выравнивание и перевод
- Семантическая эквивалентность понятий
- Модели выравнивания в машинном переводе
- False friends и межъязыковая омонимия. Автоматическое выявление ложных друзей (false friends)
- Сопоставление понятийных сетей (WordNet, BabelNet)
- Автоматический анализ лексических различий
- Генерация и анализ частотных таблиц
- Метрические методы сопоставления языков
- Машинный перевод как метод типологического анализа.
- Лингвистическая разметка и представление данных в NLP
- Стандарты лингвистической разметки
- Основные стандарты: CoNLL-U, TEI, XML
- UD (Universal Dependencies) и их применение
- Многоуровневая разметка: морфология, синтаксис, семантика
- Проблемы интероперабельности форматов
- Инструменты для чтения и конвертации форматов
- Методы автоматического получения разметки и обучение на разметке.
- Роль лингвистической разметки в обработке текста.
- Создание и использование аннотированных корпусов.
- POS-теггинг: теория и практика
- Частеречная разметка, её значение и применение
- Обзор существующих POS-теггеров (NLTK, spaCy, Stanza)
- Обучение теггеров на новых (нестандартных) корпусах
- Разработка и сравнение POS-теггеров.
- Расширенные модели: BERT и context-aware tagging

- Проблемы POS-разметки в редких языках и диалектах
- POS-теггинг для полисинтетических языков.
- Ошибки POS-разметки и способы их коррекции.
- Графовые и визуальные представления разметки
- Использование графов зависимостей (dependency parsing)
- Визуализация синтаксической структуры
- Semantic Role Labeling и графы предикат-аргумент
- Работа с библиотеками networkx, displacy
- Графовые базы данных для хранения разметки
- Математические модели в лингвистике: теория и практика
- Основы математического моделирования языка
- Вероятностные языковые модели
- Марковские цепи и HMM в NLP
- Логико-алгебраические подходы в формальной лингвистике
- Грамматики в математическом представлении
- Моделирование лексической и синтаксической структуры языка.
- Оценка и тестирование математических моделей в NLP.
- Математика и машинный перевод
- Модели выравнивания (IBM Models, Attention)
- Seq2Seq и трансформеры
- Функции потерь и оптимизация в обучении перевода
- BLEU и другие метрики качества перевода
- Машинный перевод как типологический эксперимент
- Применение математических моделей для автоматического перевода.
- Вычислительная лингвистика и большие данные
- Взаимосвязь между вычислениями и анализом естественного языка.
- Работа с корпусами в рамках Data Science
- Регрессия, кластеризация, деревья решений в текстовом анализе
- PCA, t-SNE и другие методы снижения размерности
- Роль предобученных моделей в обработке больших текстов
- Применение математических подходов для анализа больших текстовых данных.
- Концептуальные связи: язык как данные
- Метафора «язык как данные»
- Специфика языковых данных: линейность, многозначность
- Как язык «вписывается» в вычислительную парадигму
- Ограничения количественных подходов
- Гибридные модели (статистика + лингвистика)
- Машинный и нейронный перевод: подходы, архитектуры, оценка
- История, типология и роль человека в машинном переводе
- Этапы развития автоматического перевода: от правил к нейросетям
- Сравнение трансферных, статистических и нейросетевых подходов
- Роль человека в процессе машинного перевода: редактирование, обучение, сопровождение систем
- Предредактирование текста: цели, методы, рекомендации
- Постредактирование: инструменты, техники, автоматизация
- Использование автоматических систем в процессе редактирования.
- Перспективы улучшения качества машинного перевода через редактирование
- Проблемы адаптации и постредактирования переведённых текстов.
- Архитектуры и технологии нейронного перевода
- Принципы работы NMT (Neural Machine Translation)

- Архитектура Transformer и её роль в прорыве переводов
- Использование LLM (например, GPT) в мультязычных задачах
- Подходы к обучению мультязычных моделей
- Кросс-лингвистический перенос знаний в NLP
- Механизм внимания и контекстное кодирование
- Платформы и инструменты нейронного перевода
- Сравнение систем: Google Translate, DeepL, Яндекс, OpenNMT. Преимущества и ограничения открытых и закрытых платформ
- Открытые библиотеки для NMT (Marian, Fairseq, OpenNMT)
- Интерфейсы и API нейропереводчиков
- Качество перевода и языки с ограниченными ресурсами
- Метрики BLEU, METEOR, TER, COMET и их интерпретация. Сравнение автоматических и человеко-ориентированных оценок
- Проблемы качества при работе с редкими языками
- Методы сбора данных для низкоресурсных языков
- Адаптация моделей для низкоресурсных языков.
- Использование параллельных и выровненных корпусов
- Адаптация моделей с помощью transfer learning
- Методика создания выборок в лингвистике. Общие принципы лингвистических выборок
- Основные требования к выборке (репрезентативность, релевантность)
- Типы выборок: случайная, стратифицированная, тематическая
- Объём выборки и его влияние на достоверность анализа
- Этические аспекты формирования выборки
- Способы извлечения выборок из корпусов
- Частотный фильтр, случайный срез, тематические подкорпуса
- Выборка для обучения и тестирования моделей
- Создание сбалансированных подвыборок по стилю и жанру
- Роль выборок в статистическом анализе текстов.
- Оценка и валидность выборочных совокупностей
- Критерии репрезентативности и полноты
- Сравнение выборки с генеральной совокупностью
- Ошибки смещения и переобучения на выборках
- Оценка статистической устойчивости результатов
- Методы кросс-валидации в корпусном анализе
- Цифровая гуманитаристика: подходы и приложения.
- История и развитие цифровой гуманитаристики.
- Цифровые методы анализа текстов: от контент-анализа до distant reading.
- Использование цифровых корпусов в гуманитарных исследованиях.
- Визуализация текстовых данных и цифровое картографирование.
- Визуализация и интерпретация результатов работы NLP-систем.
- Примеры цифровых гуманитарных проектов (платформы, базы, исследования).
- Методы качественного анализа текста и дискурса в цифровой гуманитаристике.
- Стратегии сбора текстового материала: кейс-ориентированные подходы.
- Контент-анализ и grounded theory в корпусных исследованиях.
- Контент-анализ: Этапы, методы и технологии.
- Основные этапы подготовки контент-анализа.
- Методы обеспечения репрезентативности в контент-анализе.
- Влияние контент-анализа на лингвистические исследования.
- Факторы контроля качества эксперимента в контент-анализе.
- Современные компьютерные технологии для контент-анализа.

- Этнографические методы и нарративный подход.
- Конверсационный анализ и анализ способов участия.
- Критический дискурс-анализ и функциональная прагматика
- Моделирование в компьютерной лингвистике: методы и подходы.
- Построение статистических моделей для анализа текста.
- Применение вероятностных и логических моделей.
- Технологии обучения без учителя в обработке текстов.
- Модели на основе скрытых Марковских процессов.
- Разработка и оценка моделей машинного перевода.
- Разработка и интеграция лингвистических модулей.
- Подходы к построению модулей и систем компьютерной лингвистики.
- Модульная структура NLP-систем: архитектуры и компоненты.
- Подходы к обучению моделей: supervised vs unsupervised.
- Взаимодействие лингвистических и статистических моделей.
- Проблемы и решения в интеграции различных методов.
- Оценка качества и производительности систем NLP.
- Структурно-вероятностные модели в лингвистических исследованиях.
- Роль вероятностных моделей в изучении языка.
- Применение структурно-вероятностных моделей для анализа текста. Основные области применения таких моделей в лингвистике.
- Проблемы и вызовы при применении структурно-вероятностных моделей.
- Преимущества и ограничения структурно-вероятностного подхода.
- Модели и методы анализа диалектов.
- Обработка фонетических и орфографических вариаций.
- Моделирование языковых изменений.
- Прогнозирование языковой эволюции.
- Автоматический анализ языковых мод и трендов
- Модель BERT: Принципы и применения.
- Основные принципы работы модели BERT.
- Использование BERT для задач классификации текста.
- Преимущества предобученных моделей BERT.
- Токенизация и представление данных в BERT.
- Оценка качества модели и метрики для обучения.
- Анализ результатов классификации с BERT.
- Морфологическая обработка текста: алгоритмы, инструменты и практика.
- Методы предобработки текста: токенизация, лемматизация, нормализация, фильтрация и регулярные выражения.
- Использование регулярных выражений для фильтрации данных.
- Сравнительный анализ инструментов морфологического анализа: pymorphy2, mystem, SpaCy, NLTK.
- Расчёт частотных характеристик: POS-частота, распределение по корпусу, визуализация (matplotlib).
- Морфологический синтез и генерация словоформ: теоретические подходы и их программная реализация.
- Создание морфологического пайплайна: от загрузки данных до кэширования результатов анализа.
- Морфологический анализ и генерация.
- Построение морфологических анализаторов для разных языков.
- Сравнение правил и статистических методов морфоанализа.
- Автоматическая генерация словоформ.
- Морфология в low-resource языках.

- Использование Finite State Transducers (FST) в морфологии
- Очистка и аннотирование текстовых данных: практики предварительной обработки.
- Синтаксический и морфологический парсинг.
- Использование регулярных выражений в задачах NLP.
- Сбор и первичная обработка текстов (OCR, Data Mining).
- Обзор инструментов автоматизации (Python, Sublime, Excel).
- Предобработка и преобразование текстовых данных.
- Методы разбивки текстов на документы и предложения.
- Преобразование текста с учетом специфики задачи.
- Очистка и нормализация данных для эффективной обработки.
- Форматирование текста и его преобразование в удобный вид для анализа.
- Типы текстовых данных и подготовка корпуса.
- Частотный анализ и тематическое моделирование.
- Удаление шумов, спецсимволов, дубликатов.
- Автоматическая разметка: POS-tagging, NER, dependency parsing.
- Работа с аномалиями, опечатками и редкими формами.
- Разрешение лексической многозначности и кореференции.
- Создание пайплайна обработки данных в spaCy/NLTK.
- Современные методы предобработки текстовых данных.
- Токенизация и нормализация: алгоритмы и инструменты.
- Лемматизация и стемминг: сравнение подходов.
- Удаление шумов: стоп-слова, пунктуация, редкие термины.
- Обработка звучащей речи и мультимодальный NLP.
- ASR-системы: распознавание речи с акцентами и шумами.
- Синтез эмоциональной речи (TTS).
- Анализ аудиоподкастов: транскрипция и суммаризация.
- Видеоаналитика: совместная обработка текста и изображений.
- Приложения для людей с ограниченными возможностями.
- Анализ звучащей речи.
- Технологии распознавания и синтеза речи.
- Архитектура TTS-систем и диалоговых интерфейсов.
- Программные решения для анализа речи (Praat, ELAN).
- Методы сегментации и аннотирования аудиоданных
- Речевая аналитика и фонетические технологии в NLP.
- Извлечение признаков из аудиосигнала: мел-частотные кепстральные коэффициенты, спектрограмма.
- Сравнение ASR-систем (Whisper, Silero, Vosk): точность распознавания, шумоустойчивость.
- Синтез речи: интонационное моделирование, voice cloning, speech-to-speech generation.
- Обработка фонетических и паралингвистических аспектов речи (темп, интонация, паузы).
- Построение пайплайнов TTS и ASR: от аудио до текстов с анализом ошибок.
- Распознавание речи (ASR) и его ошибки.
- Трансформация текста в речь (TTS).
- Выделение просодических признаков (тон, ударение, паузы).
- Мультязычные модели ASR.
- Транскрибирование исторических или редких аудиозаписей
- Построение семантических сетей (WordNet, ConceptNet).
- Семантический парсинг: подходы и инструменты.
- Представление значений слов в эмбедингах.

- Семантическая неоднозначность и её автоматическое разрешение.
- Семантическое представление значений и векторизация текста.
- Эмбединги слов и документов: word2vec, GloVe, fastText, doc2vec, word2phrases: сравнение подходов.
- Вычисление семантической близости и определение синонимии: cosine similarity, T-score, MI3.
- Обработка многозначности: word sense disambiguation с использованием WordNet и алгоритмов WSD.
- Построение графов знаний, семантических сетей и тезаурусов: OWL, RDF, Neo4j.
- Визуализация семантических пространств: t-SNE, PCA, UMAP для анализа эмбедингов.
- Синтаксически-информированные эмбединги.
- Контекстуальные эмбединги (ELMo, BERT, GPT).
- Обнаружение аналогий и антонимии
- Построение зависимостной и фразовой грамматики.
- Парсеры: ручные, ML-основанные, нейросетевые.
- Universal Dependencies и их применение.
- Автоматический разбор предложений со сложной вложенностью.
- Dependency и constituency parsing: теория, инструменты, визуализация зависимостей.
- Построение синтаксических паттернов и их извлечение с помощью Rule-based и ML-инструментов.
- Сравнение парсеров (UDPipe, Tomita, Yargy, Natasha): точность, устойчивость, применимость.
- Sentence boundary disambiguation и автоматическое разделение текста на синтаксические единицы.
- Конструирование и тестирование контекстно-свободных грамматик: синтаксический разбор с помощью NLTK и Ruta
- Лингвистическая экспертиза и автоматический анализ авторства.
- Методы авторской атрибуции и стилометрии.
- Лингвокриминалистика и речевая экспертиза.
- Моделирование речевого воздействия и манипуляции.
- Стилизация текста под определённого автора.
- Обнаружение индивидуальных лингвистических маркеров.
- Перевод текста в «стиль» заданного писателя.
- Сравнительная стилометрия. ML-модели для стилистической трансформации.
- Распознавание подделок и подражаний.
- Стилометрический анализ.
- Методы извлечения стилометрических признаков текста.
- Авторство и жанровая принадлежность: стилометрия на практике.
- Выделение цитат и аллюзий. Автоматическое сопоставление с источником.
- Обнаружение и анализ плагиата.
- Алгоритмы нахождения текстовых заимствований.
- Семантический и лексический плагиат.
- Проверка переформулированных фрагментов.
- Статистика и визуализация схожести текстов.
- Мультимодальный анализ (текст + изображение/видео).
- Распознавание смыслов в подписи и изображении.
- Генерация описания изображений (image captioning).
- Соответствие эмоций текста и визуала.
- Мультимодальная риторика (реклама, мемы).

- Использование CLIP, Flamingo и других моделей
- Нейминг и брендинг в NLP.
- Анализ звучания и семантики названий. Генерация названий с помощью языковых моделей. Оценка оригинальности и узнаваемости. Сравнение с реестрами торговых марок.
- Обнаружение повторов, тавтологий и клише.
- Автоматическое выявление шаблонных фраз.
- Сравнение с базами клише и штампов.
- Повторы в студенческих и официальных текстах.
- Инструменты редактирования и улучшения стиля
- Обнаружение сарказма и иронии.
- Обучение моделей на иронических текстах.
- Сложности распознавания сарказма в соцсетях.
- Обработка данных с использованием ETL-процессов.
- Разработка ETL-пайплайнов для обработки текстовых данных.
- Интеграция данных из разных источников с использованием ETL.
- Очистка и нормализация данных в процессе ETL.
- Использование инструментов для автоматизации ETL.
- Применение ETL для подготовки данных для анализа.
- Обработка текстовых данных с использованием инструментов Big Data: Hadoop и Spark.
- Введение в обработку текстовых данных с использованием Hadoop.
- Применение Spark для работы с большими объемами текстов.
- Разработка и настройка кластеров для обработки данных.
- Оптимизация обработки текстов с помощью распределенных вычислений.
- Сравнение Hadoop и Spark для анализа текстовых данных.
- Обработка научных текстов. Распознавание терминов и понятий.
- Построение онтологий и концептографов.
- Стандартизация и нормализация формулировок.
- Автоматическое выделение результатов и гипотез.
- Обработка новостных текстов. Классификация по жанрам и темам. Выявление новостной предвзятости.
- Автоматическое составление сводок.
- Детектирование повторов и копипасты.
- Отслеживание развития событий.
- Обнаружение фейков, пропаганды и дезинформации.
- Использование NLP-моделей для фактчекинга
- Морфологический, синтаксический и семантический анализ.
- Системы генерации и анализа текстов на естественном языке.
- Классификация текстов с применением алгоритмов машинного обучения.
- Оптимизация структуры данных и форматирование текстов.
- Использование кодировок и преобразование форматов.
- Извлечение данных из Twitter, Reddit, Telegram.
- Анализ хештегов и эмодзи как маркеров мнений.
- Выявление ботов и фейковых аккаунтов.
- Обработка разговорного и жаргонного языка.
- Прикладная лингвистика в цифровую эпоху: от теории к практике.
- Эволюция понятий и задач прикладной лингвистики.
- Методы анализа и моделирования языковой информации.
- Технологии представления знаний: гипертекст и онтологии
- Применение ChatGPT и LLM в гуманитарных исследованиях.

- Генерация гипотез и аналитических обзоров.
- Эвристические задачи в стилистике и поэтике.
- Критика и ограничение LLM в интерпретации.
- Интеграция моделей в учебный процесс и DH-проекты
- Программы и платформы для Text Mining.
- GATE: архитектура, компоненты, применение.
- Orange и RapidMiner в задачах классификации и анализа.
- Opinion mining в социальных медиа и новостях.
- Автоматическое обнаружение фейков, хейта и токсичности.
- Создание собственных NLP-пайплайнов без программирования
- Проект Natasha: Инструменты и библиотеки для русского NLP.
- Обзор библиотеки Natasha для обработки русского языка.
- Разработка и использование модели NER для русского языка.
- Применение Razdel для сегментации русского текста.
- Использование Slovnet для глубокого анализа русского языка.
- Работа с Naves для создания эмбедингов русского языка.
- Использование Nerus для создания синтетического датасета с разметкой.
- Промпт-инженерия и оптимизация работы с LLM.
- Дизайн эффективных промптов для разных задач. Few-shot и zero-shot обучение: стратегии и примеры.
- Сравнение моделей (GPT-4, Claude, Mistral) в контексте запросов.
- Автоматизация тестирования и улучшения промптов.
- Безопасность: предотвращение jailbreak-атак и misuse.
- Базовые принципы создания эффективных промптов.
- Шаблоны промптов: zero-shot, few-shot, chain-of-thought.
- Контекстуализация запросов: установка ролей, ограничение задач.
- Отладка и итеративное улучшение промптов.
- Примеры промпт-инжиниринга для гуманитарных задач
- Работа с ИИ-текстами: постредактирование и улучшение.
- Критика и редактирование сгенерированного текста.
- Улучшение текстов по параметрам жанра, стиля и цели. Проверка грамматики и логической согласованности.
- Автоматизированные инструменты оценки и коррекции текста
- Работа с рукописями и OCR.
- Компьютерная текстология и анализ древних рукописей.
- Распознавание текста на изображениях.
- Очистка результатов OCR и нормализация. Сопоставление сканов с цифровыми текстами
- Работа с нестандартной типографикой. Распознавание исторических шрифтов.
- Разработка API и микросервисов для обработки текста с помощью FastAPI.
- Основы FastAPI: структура, эндпоинты, сериализация.
- Реализация простых NLP-инструментов как REST-сервисов.
- Обработка запросов: асинхронность и масштабируемость.
- Интеграция с моделями машинного обучения.
- Деплой и документация через Swagger / ReDoc.
- Семантические сети и онтологии.
- Построение онтологий вручную и автоматически.
- Связывание сущностей и отношений.
- Использование Wikidata, DBpedia, YAGO.
- Интерфейсные решения для работы с онтологиями.
- Проверка согласованности и полноты онтологий

- Сложности в компьютерной обработке естественного языка.
- Проблемы многозначности слов и контекста.
- Трудности в понимании и генерации естественного языка.
- Языковые особенности и их влияние на обработку текста.
- Как модели NLP решают проблемы семантики и синтаксиса.
- Стратегии улучшения точности обработки языка.
- Проблемы и вызовы в автоматической обработке текста.
- Ошибки и искажения в распознавании текста.
- Проблемы качества разметки и меток данных.
- Ограничения и вызовы в обработке слабо ресурсоемких языков.
- Этические аспекты и проблемы конфиденциальности данных
- Создание пайплайнов для обработки данных с Apache Airflow.
- Основы Apache Airflow и его роль в обработке данных.
- Разработка пайплайнов для обработки текстовых данных.
- Использование Airflow для автоматизации ETL-процессов.
- Интеграция Airflow с другими инструментами для анализа текстов.
- Применение Airflow в реальных задачах обработки данных
- Создание скриптов на VBA для автоматизации анализа текста.
- Разработка макросов для извлечения данных из текстовых файлов.
- Построение скриптов для автоматической очистки и подготовки данных.
- Реализация алгоритмов для текстовой кластеризации в Excel.
- Автоматизация анализа частотности слов и фраз.
- Применение регулярных выражений для поиска паттернов в тексте.
- Сравнение языковых моделей и инструментов.
- Тестирование моделей на одинаковых задачах. Оценка скорости, точности, адаптивности. Примеры применения: spaCy, Stanza, BERT, ChatGPT.
- Условия обучения и качество данных.
- Разработка мини-бенчмарков для лингвистов
- Статистический анализ текстов с использованием SPSS и Jamovi.
- Основы статистического анализа текста в SPSS.
- Применение Jamovi для обработки текстовых данных.
- Оценка частотности и распределения слов с помощью статистики.
- Модели для проверки гипотез на основе текстовых данных.
- Сравнение SPSS и Jamovi для анализа текстов.
- Текстовые модели в психолингвистике.
- Определение тревожности, депрессии, агрессии по тексту.
- Анонимизация чувствительной информации.
- Корпуса психолингвистических текстов.
- Психолингвистический анализ текста (LIWC и др.).
- Тематическое моделирование.
- LDA, NMF, LSI, BERTopic и другие тематические модели.
- Выявление скрытых тем в корпусах.
- Классификация новостей и постов.
- Динамика тем во времени.
- Визуализация тематических карт.
- Тематическое моделирование с использованием Gensim.
- Основы тематического моделирования и его применение.
- Оценка и улучшение качества тематического моделирования.
- Применение модели для анализа больших объемов текстовых данных.
- Визуализация результатов тематического моделирования.

- Этические и правовые аспекты NLP и ИИ.
- Этические проблемы генерации текста (фейк-ньюс, плагиат и др.).
- Прозрачность и интерпретируемость языковых моделей.
- Регулирование и стандарты ИИ в образовании и науке.
- Авторское право и ИИ-контент: текущее состояние и дискуссии.
- Генерация токсичного или фейкового контента.
- Проблемы приватности при анализе текста.
- Этические ограничения при сборе корпусов.
- Объективность и предвзятость моделей.
- Юридические аспекты генерации контента
- Компьютерная лингвистика и искусственный интеллект: пересечение и развитие технологий.
- Основные этапы становления компьютерной лингвистики.
- Влияние искусственного интеллекта на развитие NLP.
- Применение методов ИИ для обработки естественного языка.
- Компьютерная лингвистика: методы, направления, вызовы.
- Междисциплинарный характер и задачи компьютерной лингвистики.
- Стадии обработки текста: от предобработки до интерпретации.
- Статистический и лингвистический подходы к построению систем.
- Модели тематического анализа и лексическая семантика
- Распознавание и группировка культурно-значимых слов.
- Лингвокультурные маркировки в текстах. Автоматизация лингвокультурного анализа.

Устный опрос:

- Причины появления компьютерной лингвистики
- Когда появилась компьютерная лингвистика?
- Назовите основные этапы развития компьютерной лингвистики.
- Какие технологические прорывы стали ключевыми для NLP?
- Дайте определение «компьютерной лингвистике».
- Как появление LLM повлияло на задачи и методы компьютерной лингвистики?
- Что такое NLP? Как соотносятся понятия NLP и компьютерной лингвистики?
- Как архитектура Transformer повлияла на NLP?
- Как компьютерная лингвистика связана с искусственным интеллектом?
- Можно ли считать синонимами прикладную и компьютерную лингвистику?
- Чем знаменит Джорджтаунский проект?
- Перечислите основные направления компьютерной лингвистики. В чем особенности каждого направления?
- Перечислите уровни организации языка. В чем практическая значимость их выделения для компьютерной лингвистики?
- Какие стратегии машинного перевода существуют?
- Сравните rule-based, статистический и нейронный машинный перевод
- Как лингвистические знания используются при информационном поиске?
- В чем отличия между аннотированием и реферированием текста?
- Какова практическая значимость реферирования?
- В чем отличия между классификацией и кластеризацией?
- Что такое Text Mining?
- Что такое Data Mining? Чем Data Mining отличается от Text Mining?

- В чем особенности вопросно-ответных систем? Отличия от ИП?
- Что подразумевает индексирование?
- Как происходит извлечение информации?
- Для чего проводится извлечение именованных сущностей?
- Что такое анализ тональности текстов? Для чего он используется?
- Для чего применяются чат-боты?
- В чем отличие чат-ботов от диалоговых систем?
- Какова роль спеллеров?
- Какие подходы к выявлению и исправлению ошибок существуют в КЛ?
- Какие способы автоматической генерации текста вам известны?
- Что такое генерация текста на основе LLM и чем она отличается от классических методов (n-граммы, шаблоны)?
- Что такое NLU? В чем сложность NLU?
- Что включают задачи распознавания и синтеза речи?
- Чем обусловлены сложности распознавания и синтеза звучащей речи?
- Какие существуют подходы к распознаванию речи?
- Чем обусловлена сложность моделирования в КЛ?
- Перечислите основные модули обработки текстов.
- Что такое морфема?
- Что такое лексема? Чем она отличается от словоформы?
- Что такое анафорические ссылки?
- Что такое кореференция?
- Какие явления неоднозначности вам известны?
- Что такое омонимия? Какие виды омонимии вы знаете? В чем значимость этой информации для КЛ?
- Перечислите этапы компьютерного анализа текстов на естественном языке? В чем особенность каждого из них?
- Что включает графематический анализ?
- Что включает морфологический анализ?
- Что включает синтаксический анализ?
- Что включает семантический анализ?
- В чем заключаются сложности выделения предложений из текста? А выделения слов?
- Что включает нормализация?
- Что такое токенизация?
- Что такое токен?
- Как токенизация учитывает особенности разных языков (например, китайского без пробелов)?
- Чем лемматизация отличается от стемминга? Когда предпочтительнее каждый метод?
- Что такое лемма?
- Что такое стоп-слова? Как они обычно обрабатываются при анализе текстов?
- Что такое сема?
- Что такое коллокация? В чем значимость этой информации для КЛ?
- Что такое N-граммы?

- Какие виды лексических ресурсов вам известны? Какие лексические ресурсы используются в NLP?
- Какие типы словарей вы знаете?
- В чем особенность частотных словарей?
- Что такое тезаурус?
- Что такое онтология?
- Что такое синсет?
- Что такое гипонимия? Что такое меронимия?
- Что такое корпус текстов? Для чего они используются в компьютерной лингвистике?
- Чем корпус отличается от библиотеки текстов?
- Какие типы корпусов вам известны?
- Как вычисляется *irm*? Для чего она используется?
- Что такое разметка?
- Какие виды разметки вы знаете? Какая разметка представлена в НКРЯ?
- Как размечаются корпуса?
- Какие современные корпуса вам известны?
- Что такое корпусный менеджер?
- Какими свойствами должен обладать корпус?
- Что подразумевает репрезентативность корпуса?
- В чем отличия между репрезентативностью и сбалансированностью?
- Из чего состоит корпус?
- Какие подходы к построению модулей и систем КЛ вы знаете?
- В чем особенности rule-based подхода? В чем преимущества и недостатки rule-based подхода?
- Как эволюционировала компьютерная лингвистика от rule-based систем к нейросетевым моделям?
- Как LLM (например, GPT) изменили подходы в NLP?
- Что такое машинное обучение?
- Какие методы машинного обучения вам известны?
- Чем обучение с учителем отличается от обучения без учителя?
- Что такое обучающая выборка? Для чего она используется? Как готовится?
- В чем преимущества и недостатки машинного обучения?
- Как происходит векторизация текстов? Для чего она необходима?
- Что представляет собой модель bag of words?
- Что представляет собой модель tf-idf? Как рассчитывается tf-idf? Для чего применяется?
- Что представляет собой модель word2vec? Для чего применяется?
- Как происходит обучение нейронных сетей?
- Что такое лингвистическая защита информации? Как лингвистические знания используются при защите информации?

ИПС:

- Что такое информационная потребность?
- Как соотносятся информационная потребность и поисковый запрос?

- Что такое коллекция документов (в ИПС)?
- Что включают технические и содержательные проблемы при разработке ИПС? Чем они отличаются?
- Для чего используется инвертированный индекс? Из чего он состоит?
- Что такое комбинаторный взрыв? Когда он возникает?
- В чем сложности разработки ИПС?
- Как принято классифицировать запросы? В чем особенности каждого типа? (информационные, навигационные, транзакционные)
- Что такое семантический индекс? Для каких типов запросов его рационально использовать?
- Что такое релевантность поиска?
- Как определить полноту поиска?
- Как определить точность поиска?
- Как происходит ранжирование поисковой выдачи?
- Как происходит фильтрация поисковой выдачи?
- Для чего применяются факторы ранжирования? Какие типы факторов ранжирования вы знаете?
- Что включают текстовые факторы ранжирования?
- Что включают ссылочные факторы ранжирования?
- Что включают поведенческие факторы ранжирования?
- Как оценивается релевантность поиска?
- Чем занимаются ассессоры?
- Для чего необходимо взвешивать термины при поиске? Как при поиске применяется tf-idf?
- Для чего в ИПС используется компьютерная лингвистика?
- Какие вы знаете стратегии обработки опечаток?
- Для чего в ИПС используются расширения терминов запроса? Как они работают?
- Как при поиске могут применяться тезаурусы?
- Что представляют собой тезаурусные расширения? Для чего используются? Как происходит сбор словарей расширений?
- В чем особенности координатного инвертированного индекса? Когда он используется?
- Как векторные модели (Word2Vec, BERT) улучшают релевантность поиска?
- Почему пертинентность сложнее измерить, чем релевантность?

Машинный перевод:

- Каковы три подхода к машинному переводу? Каковы особенности каждого из них?
- Когда и как возник машинный перевод?
- В чем отличия машинного перевода от дешифровки?
- Чем машинный перевод отличается от автоматизированного?
- Какие существуют типы перевода на основе правил? В чем особенности каждого из них? (трансферная, пословная, интерлингвистическая)
- Что такое интерлингва?
- Каковы этапы трансферного машинного перевода?
- Что такое база памяти переводов? Для чего она используется?
- Как работают статистические системы машинного перевода?

- Чем нейронный машинный перевод (NMT) лучше статистического?
- Для чего в статистическом переводе используются модель языка и модель перевода? Что они собой представляют?
- Как строится модель языка? Какие есть проблемы при ее построении? Как они решаются?
- Для чего используются цепи Маркова? Как они работают? Какие типы цепей Маркова бывают? Какие чаще всего используются?
- Что включает триграммная языковая модель? Как применяется?
- Как происходит оценка максимального правдоподобия? Для чего она необходима?
- Какие методы сглаживания вам известны? Для чего они применяются? (линейная интерполяция, методы дисконтирования,
- Для чего применяется модель перевода?
- Для чего в машинном переводе применяются корпуса? Какие типы корпусов чаще всего используются?
- Что такое выравнивание? Для чего оно необходимо? Какие сложности могут возникнуть при выравнивании?
- Как происходит автоматическое выравнивание?
- Как строится матрица выравнивания? Для чего она используется?
- Как выравнивание корпусов используется в современных системах (например, OPUS)?
- Что такое консистентность? Как можно ее проверить при помощи матрицы выравнивания?
- Каковы этапы перевода на основе сочетаний?
- Какие подходы к гибриднему переводу существуют? Когда каждый из них целесообразно использовать?
- Какие методы оценки качества перевода существуют?
- Как происходит автоматическая оценка перевода? Приведите примеры метрик
- Для чего используются CAT-tools? Приведите примеры таких систем.
- Как механизм внимания в трансформерах улучшил качество перевода?

Чат-боты и диалоговые системы:

- В чем отличия диалоговых систем и чат-ботов?
- В чем заключается тест Тьюринга? Для чего он используется?
- Для чего была учреждена премия Лёбнера?
- Когда создали первый чат-бот? Каковы были его функции? Как он работал?
- Каково практическое применение чат-ботов?
- Как работают чат-боты на шаблонах? Каковы их преимущества и ограничения?
- Каковы особенности диалога на естественном языке? Как эти знания применяются в КЛ?
- В чем сложности создания диалоговой системы, полностью отвечающей особенностям диалога на естественном языке?
- Какова архитектура диалоговых систем? Какие элементы являются обязательными?
- В чем задача модуля понимания естественного языка? Какие существуют подходы к его разработке?
- С какой целью в диалоговых системах используются базы знаний? Как они формируются?

- Что такое диалоговый менеджер? В чем его задачи?
- Из каких стадий состоит процесс генерации языка? Опишите каждую стадию.
- Как работают чат-боты? Перечислите этапы.
- Для чего применяется язык AIML? Назовите основные теги этого языка?
- Какие ресурсы позволяют создавать чат-ботов?
- Для чего может использоваться JAICP?
- Каковы функции Dialogflow CX и Watson? В чем особенность каждого?
- Как работает модель последовательность к последовательности (seq2seq)? Для чего она используется в диалоговых системах?
- Каковы перспективы углубления диалога чат-ботами и диалоговыми системами?
- В чем разница между диалоговыми системами на правилах и на LLM?

Генерация текста:

- В чем отличие между генерацией текста и перефразированием?
- Когда была создана первая система генерации текста? Каковы были её возможности?
- Какие этапы развития прошли модели генерации текста?
- В чем разница между правилами, шаблонами и нейросетевыми моделями в генерации текста?
- Для чего используются синонимайзеры? Какие есть виды синонимизации?
- Как генерация текста используется в чат-ботах и виртуальных ассистентах?
- Каковы этические риски автоматической генерации текста?
- Как работают n-граммные модели в генерации текста? Каковы их ограничения?
- Что такое марковские цепи и как они применяются для генерации текста?
- Как рекуррентные нейронные сети (RNN, LSTM) используются для генерации текста?
- В чем преимущество трансформеров перед RNN в генерации текста?
- Как работает модель seq2seq в генерации текста? Где она применяется?
- Что такое авторегрессионные модели (например, GPT)? Как они генерируют текст?
- Как VAE (вариационные автоэнкодеры) и GAN (генеративно-состязательные сети) применяются в генерации текста?
- Какие метрики (BLEU, ROUGE, Perplexity) лучше использовать для оценки генерации текста?

LLMs:

- Что такое AGI? Как он соотносится со слабым ИИ?
- Что такое искусственные нейронные сети? Каков принцип их работы?
- Что такое LLM? Почему они так называются?
- Что такое генеративные языковые модели? Назовите примеры. Почему они так стремительно развиваются?
- В чем особенности ChatGPT? Каковы возможности и ограничения?
- Как работает глубокое обучение?
- Что такое эмбединги?
- Что такое матрица эмбедингов?
- В чем особенность модели трансформеров?
- Что такое энкодер и декодер?

- Что такое механизм внимания?
- Как устроен механизм self-attention в Transformer?
- Как моделируется язык?
- Что происходит на этапе pre-training?
- Как происходит подготовка данных для обучения? Как фильтруют данные?
- Что такое токены? Как в LLM происходит токенизация? Назовите проблемы токенизации.
- Что такое алгоритм BPE?
- Почему BERT использует subword-токенизацию, а GPT — байтовые пары (BPE)?
- Как происходит обучение нейросети?
- Как происходит векторизация токенов?
- Что такое контекстное представление?
- Как модель предсказывает следующий токен?
- Что происходит на этапе post-training? Назовите два этапа пост-тренировки.
- Что такое supervised fine-tuning?
- Что такое обучение с подкреплением?
- Что такое RLHF? Как ChatGPT использует RLHF для улучшения диалога?
- Что происходит на этапе инференса?
- Что такое GPT? Как она устроена и в чем её революционность?
- Как устроена архитектура GPT? Какие этапы обучения (pre-training, fine-tuning) она включает?
- Какова история GPT?
- Как BERT отличается от GPT в плане генерации текста?
- Что такое T5, BART и другие трансформерные модели? Как они генерируют текст?
- Как модель выбирает следующее слово? Назовите 3 подхода.
- Что такое zero-shot и few-shot learning? Приведите примеры.
- Как few-shot и zero-shot обучение позволяют моделям генерировать текст без тонкой настройки?
- Как работает few-shot learning в GPT-4?
- Что такое temperature, top-k и top-p сэмплирование? Как они влияют на качество текста?
- Как работает beam search в генерации текста? Каковы его плюсы и минусы?
- Что такое промпт-инженерия? Как правильно формулировать запросы для генерации текста?
- Как prompt engineering влияет на качество ответов LLM?
- Как использовать промпт-инженерию для улучшения ответов ChatGPT?
- В чем разница между fine-tuning и prompt engineering для LLM?
- Что такое zero-shot prompting, one-shot prompting, few-shot prompting, chain-of-thought?
- Каковы общие рекомендации по использованию ChatGPT-like систем?
- Как контролировать стиль, тональность и содержание сгенерированного текста?
- Какие метрики используются для оценки качества сгенерированного текста (BLEU, ROUGE, Perplexity)?
- В чем проблема повторов и противоречий в сгенерированном тексте? Как её решают?
- Как когерентность и связность текста влияют на его восприятие?

- Что такое галлюцинации ИИ в генерации текста? Почему модели галлюцинируют?
- Как галлюцинации ИИ влияют на доверие к моделям? Какие методы их минимизируют?
- Какие современные методы борьбы с галлюцинациями LLM существуют?
- Что такое рабочая память модели? На что она влияет?
- Что такое мультимодальные языковые модели? Как они используются в компьютерной лингвистике?
- Как Hugging Face Transformers упрощает работу с NLP-моделями?
- Какие этические риски связаны с генеративным ИИ (дезинформация, deepfake)?
- Каковы основные вызовы этики, безопасности и социальной ответственности при использовании LLM?
- Что такое Retrieval-Augmented Generation (RAG) и как этот подход меняет задачи извлечения информации?

Машинное обучение:

- Опишите задачу машинного обучения.
- Что такое множество объектов? Множество ответов? Целевая функция? Приведите пример.
- Что такое кластер?
- Что такое функционал качества?
- Какие шаги необходимо выполнить для реализации машинного обучения?
- Какие группы методов машинного обучения вы знаете? В чем особенности каждого из них?
- Опишите метод ближайших соседей.
- Опишите Метод k-ближайших соседей.
- Опишите Метод парзеневского окна.
- Опишите Наивный байесовский классификатор.
- Опишите Метод опорных векторов.
- Опишите Искусственные нейронные сети.
- Что такое обратная связь?
- Что такое выбросы?
- Какие группы алгоритмов кластеризации вы знаете?
- Каковы особенности графических алгоритмов кластеризации? В чем их преимущества?
- Какие графические алгоритмы кластеризации вы знаете?
- Каковы особенности статистических алгоритмов кластеризации?
- Как работает EM-алгоритм?
- Каковы особенности иерархических алгоритмов кластеризации?
- Для решения каких задач компьютерной лингвистики используется кластеризация?
- Как выбирать метод машинного обучения для решения конкретной задачи компьютерной лингвистики?
- Приведите классификацию возможных ответов алгоритма (которая зависит от ожидаемых и полученных результатов классификации).
- Отобразите матрицу ошибок.
- Как рассчитывается метрика Accuracy, и что она оценивает?
- Как рассчитывается метрика Precision, и что она оценивает?

- Как рассчитывается метрика Recall, и что она оценивает?
- Как соотносятся метрики Recall и Precision?
- Для чего используется F-Score? Как она рассчитывается, и что оценивает?
- Как рекомендуется рассчитывать метрики, если у нас несколько классов?
- В чем состоит подход macro-averaging?
- Для чего используется ROC-кривая? Как она строится? Как ее интерпретировать?
- Что отражает AUC?
- В каких случаях рекомендуется использовать каждую из рассматриваемых метрик?
- Что представляет собой терм-документная матрица? Для чего она используется?

Анализ тональностей:

- Чем занимаются системы извлечения мнений?
- Какую информацию позволяет извлечь анализ тональности?
- В каких типах текстов обычно содержится эмоциональная оценка?
- С чего начинается анализ тональности?
- Что такое UGC? В чем его особенности? Как они влияют на обработку?
- Какие составляющие мы выявляем при анализе тональностей?
- Сколько классов выделяют при анализе тональностей?
- Как количество классов влияет на точность классификации?
- С какими сложностями можно столкнуться при анализе тональностей?
- Как измерить тональность текста?
- Какие две большие группы методов выделяются в компьютерной лингвистике? В чем их преимущества и недостатки? Когда используется каждый из них?
- Как строятся правила для анализа тональностей?
- Какие сложности возникают при построении правил для анализа тональностей?
- Как рассчитывают итоговую окраску всего документа?
- Как строятся словари оценочной лексики? Какая информация в них может быть отражена?
- Как создаются словари оценочной лексики?
- Как можно расширять списки слов в словарях оценочной лексики? (добавлять новые слова)
- Как рассчитывается PMI? Для чего используется? Как рассчитывается SO?
- Как рассчитывается индекс нейтральности/объективности?
- Как при анализе тональностей использовать методы машинного обучения?
- Как размечается обучающая выборка для анализа тональностей?
- Что такое «золотой стандарт»? Для чего он используется?
- В чем практическая значимость анализа тональностей?
- Как трансформеры (например, RoBERTa) улучшили анализ тональности?

Практика:

- Что такое OCR?
- Охарактеризуйте основные возможности OCR-программ. Каковы перспективы развития OCR-программ?
- Какие поисковые операторы вы знаете? Для чего они применяются?

- Что такое формальная и смысловая релевантность поиска? Как различие этих понятий отражается на результатах поиска?
- Как вы понимаете пертинентность? Какие способы снижения пертинентности вы можете предложить?
- Охарактеризуйте два основных типа информационно-поисковых систем: документальные и фактографические.
- В чем состоят различия информационно-поисковых систем с ручным и автоматическим индексированием? Приведите примеры систем обоих типов.
- Что такое общий и специализированный каталог веб-ресурсов? Приведите примеры каталогов обоих типов.
- Что такое фасетная классификация? Приведите примеры фасетов при описании одного документа.
- Каков функционал Voyant Tools? Для чего его можно использовать?
- Какие инструменты позволяют анализировать тренды и частотность слов?
- Какие инструменты позволяют анализировать коллокации?
- Для чего применяется Palladio? Каков его функционал?
- Для чего применяется OpenRefine? Каков его функционал?
- Какие инструменты позволяют очищать наборы данных?
- Для чего применяется Rapid Miner? Каков его функционал?
- Что такое парсинг? Какие этапы он включает?
- Для чего могут применяться ParseHub и Octoparse? В чем их особенности?
- Какие инструменты могут использоваться для парсинга? В чем их преимущества, особенности и недостатки?
- Для чего используются морфологические анализаторы? Приведите примеры морфологических анализаторов.
- Для чего применяется Mystem? Каков его функционал?
- Какой словарь используется для морфологической разметки русского языка?
- Какие библиотеки Python используются для решения задач компьютерной лингвистики?
- Какие библиотеки Python (Hugging Face, spaCy) наиболее эффективны для NLP?
- Какими способами можно производить токенизацию в Python? В чем отличия разных способов токенизации?
- Как можно проводить стемминг и лемматизацию в Python? В чем отличия результатов после стемминга и лемматизации?
- Как можно очищать текст от пунктуации и стоп-слов в Python?
- Как можно посчитать количество слов в Python?
- Какие библиотеки Python можно использовать для морфологического анализа? Каковы особенности каждой библиотеки?
- Какие библиотеки Python можно использовать для морфологического синтеза? Для чего может быть полезна эта задача?
- Для чего можно использовать методы similar, common_contexts и collocations из библиотеки NLTK?
- Как и для чего проводится кэширование лемм?
- Какие библиотеки Python позволяют работать с векторными моделями? Перечислите их основные методы и для чего они могут применяться.
- В чем отличия word2vec, Glove и fastText?
- Какие библиотеки Python позволяют извлекать именованные сущности?

- Каковы цели тематического моделирования? Для решения каких задач оно применяется? Какие программы позволяют его провести?
- Для чего применяется Topic Modeling Tool?
- Что такое стоп-слова? Как они обычно обрабатываются при анализе текстов?
- Какая предобработка данных проводится при тематическом моделировании?
- Какие функции табличных редакторов могут быть полезны при анализе текстовых данных?
- Для чего используется визуализация данных в компьютерной лингвистике?
- Что такое Distant reading? Чем он отличается от Close reading?
- Что такое Digital humanities?
- Для чего используется сетевой анализ? Каковы его преимущества? При помощи каких программ можно его проводить?
- Для чего применяется Gephi?
- Как Gephi помогает анализировать сети коллокаций?
- Для чего применяется Easy Linavis?
- Что такое граф? Что такое мультиграф?
- Как облака слов и сетевые графы помогают анализировать текстовые данные?
- Из чего состоит граф?
- Что такое сообщество в сетевом анализе?
- Какие бывают графы? (ориентированные\неориентированные, связные\несвязные, взвешенные\невзвешенные, мультимодальный)
- Какие метрики используются для анализа графа?
- Что такое степень узла? Что такое взвешенная степень?
- Какими способами можно оценить важность узла? (центральность по степени, по собственному вектору, по близости, по посредничеству, пр.) В чем особенность каждого?
- Что такое укладка графа?
- Как рассчитываются: Коэффициент ассортативности, Коэффициент кластеризации, Плотность графа, Модулярность?
- Как определяется вес ребра? Размер сети? Плотность сети? Длина пути? Клика?
- В чем отличие сложных сетей от случайных графов Эрдеша-Реньи?
- Как степенной закон распределения связей влияет на свойства безмасштабных сетей?
- Какие два условия в модели Барабаши-Альберт приводят к формированию безмасштабных сетей?
- Почему социальные сети обладают высокой кластеризацией? Приведите пример.
- Как феномен «тесного мира» проявляется в эксперименте Милграма?
- Какие методы используются для выявления сообществ в сетях?
- Почему безмасштабные сети устойчивы к случайным атакам, но уязвимы к целенаправленным?
- Как синхронизация в сетях зависит от их топологии?
- Что такое каскадные процессы в сетях? Приведите пример из реальной жизни.
- Как матрица смежности и Лапласиан применяются для анализа сетей?
- Что такое степень узла (degree)? Коэффициент кластеризации? Ассортативное смешивание? Перколяция?

- Каким образом можно визуализировать текстовые данные? Назовите примеры программ, позволяющих визуализировать текстовые данные.
- Для чего используются облака слов? При помощи каких ресурсов их можно построить?
- Для чего применяется Tableau?
- Для чего можно использовать WordNet? Какой у него функционал?
- Что такое стилеметрия? Какие инструменты можно использовать для ее проведения?
- В чем особенности метода Delta?
- Что такое атрибуция текста?
- Какие задачи позволяет решать стилеметрия?
- Что анализирует стилеметрия? (Что считают алгоритмы? На основании чего алгоритм делает выводы?)
- Какая метрика в стилеметрии наиболее распространенная?
- Для чего применяется функция `oppose` в R?
- Перечислите наиболее важные задачи речевой обработки и анализа. Какова практическая значимость решения этих задач?
- Что такое регулярные выражения? Для чего они используются?
- Приведите примеры регулярных выражений. Как происходит группировка символов?
- Какие регулярные выражения эффективны для обработки текста? (примеры: извлечение email, хештегов)
- Что такое «жадные» и «ленивые» квантификаторы?
- Для чего применяется экранирование?
- Что такое язык разметки?
- Что такое TEI? Из чего состоит стандартный документ TEI?
- Что включают в себя именованные сущности? Для чего их выделяют?
- Для чего используется BaseX?
- Для чего используется язык запросов XPath? Каковы особенности его синтаксиса?
- Что такое файловая система? Как она устроена? Что такое FAT? Из каких частей она состоит?
- Что такое фрагментация и дефрагментация?
- Что такое кодировка? Какие проблемы могут возникать из-за кодировок?
- Расскажите краткую историю кодировок (ASCII, ISO/IEC 646, KOI8, ISO/IEC 8859, Windows-125x, UTF-8, Unicode)
- Как расширение файла связано с кодировкой? В чем преимущества .pdf?
- Что такое макросы?
- Какие ресурсы для проведения Text Mining вы знаете? Каков их функционал?
- Приведите примеры анализаторов текста. Какой их функционал? Каковы возможности анализа?
- Какие ресурсы позволяют проводить анализ частотности упоминаний? Совместной встречаемости? Анализ сюжета? Анализ идиостиля писателя?
- Какие инструменты позволяют автоматически определять функциональный стиль текста? Проводить синтаксический анализ? Проводить анализ морфологии?
- Назовите и кратко охарактеризуйте уровни естественного языка, релевантные для морфологического анализа и синтеза текста.
- Назовите и дайте краткую характеристику этапам автоматического анализа текста.

- Назовите и дайте краткую характеристику этапам автоматического синтеза текста.
- Охарактеризуйте системы компьютерного моделирования диалогов, в том числе роботы-автоответчики. Как происходит обучение роботов? Как распознать робот-автоответчик?
- Для чего применяется Томита-парсер? Из каких лингвистических процессоров он состоит? Каковы его основные компоненты? Каков принцип работы?
- С какими сложностями можно столкнуться при решении задач компьютерной фонетики?
- В каких сферах используется распознавание говорящего по голосу?
- Чем верификация отличается от идентификации?
- Перечислите этапы процесса распознавания говорящего.
- Для решения каких задач применяется распознавание речи (STT)?
- Что такое коллокации? Что такое N-gramm? Для чего они используются в компьютерной фонетике?
- Перечислите этапы преобразования речи в текст.
- Для чего используется Praat?
- Для чего используется ELAN?
- Что такое спектрограммы? Для чего они используются? Как их можно «прочитать»?
- Чем осциллограммы отличаются от спектрограмм?
- Что такое форманты? Для чего они используются в компьютерной фонетике?
- Перечислите спектральные характеристики речевого сигнала.
- Для чего проводится анализ спектральных характеристик речевого сигнала? При помощи каких программ его можно проводить?
- Перечислите критерии для оценки качества синтеза речи.
- Перечислите типы систем TTS. Чем они отличаются?
- Что такое дифоны? Для чего они используются?
- Какие инструменты можно использовать для реализации статистического синтезатора речи?

Квантитативная лингвистика:

- Чем занимается квантитативная лингвистика? Какие прикладные задачи она помогает решить?
- Как квантитативная лингвистика может применяться при дешифровке? Приведите примеры. Почему это все еще актуально?
- Что такое дистрибутивная семантика?
- Что такое типология? Как методы квантитативной лингвистики могут использоваться в типологии?
- Как можно измерить сложность языков количественными методами?
- Что такое стилеметрия?
- Каковы практические применения стилеметрии? Приведите примеры.
- Какие лингвистические параметры считают в стилеметрических исследованиях?
- Какие методы применяют в стилеметрических исследованиях?
- Какой метод наиболее часто используется в стилеметрических исследованиях?
- Перечислите этапы стилеметрического исследования.
- Как методы квантитативной лингвистики применяются для определения возраста языков?

- Что такое глоттохронология?
- В чем особенность частотных словарей? Для чего они применяются? В чем сложность создания? Каковы ограничения?
- Что такое словник?
- В чем заключается закон Ципфа?
- Как закон Ципфа объясняет частотность слов в корпусах?
- Для чего используется и что отражает показатель R (range)?
- Для чего используется и что отражает коэффициент Жуйана?
- Что такое генеральная совокупность и выборочная совокупность? В чем сложность их создания для частотного словаря?
- Для чего используется и что отражает коэффициент логарифмического правдоподобия (LL-score)?
- Чем занимается квантитативная морфология? В чем практическая значимость?

Корпусная лингвистика:

- Чем занимается корпусная лингвистика?
- Когда появился первый корпус? Каковы были его особенности?
- Для чего корпуса применяются в современной лингвистике?
- Перечислите основные свойства корпуса.
- Что такое корпусный менеджер?
- Что такое конкорданс? Для чего он применяется?
- В чем преимущества формата KWIC?
- Что такое репрезентативность? В чем сложность создания репрезентативного корпуса?
- В чем состоят проблема объема и проблема отбора при создании корпуса?
- Что такое сбалансированность?
- Почему репрезентативность и сбалансированность важны при составлении корпуса?
- Что такое разметка? Для чего она необходима?
- Что такое аннотирование?
- Какие бывают корпуса?
- Что представляют собой мультимодальные корпуса?
- Что представляют собой параллельные корпуса? Приведите примеры параллельных корпусов.
- Что такое выравнивание? Какие два способа выравнивания существуют?
- Как происходит автоматическое выравнивание? Какие методы автоматического выравнивания существуют? Назовите программы для автоматического выравнивания.
- Для чего используется ABBYY Aligner?
- Какие бывают корпуса по критерию параллельности? Для чего они используются?
- Какие бывают корпуса по типу языковых данных?
- В чем сложность создания репрезентативного корпуса устной речи?
- Перечислите типы разметки.
- Для чего используются теггеры и парсеры?
- Какие типы программ разметки существуют?
- Какие типы лингвистической разметки существуют?

- Каковы принципы создания лингвистической разметки?
- Что такое POS-tagging?
- Что включает в себя морфологическая разметка?
- Что включает в себя семантическая разметка?
- Что включает в себя синтаксическая разметка?
- Что включает в себя метаразметка?
- В чем заключается задача разрешения омонимии?
- В чем особенности НКРЯ?
- Какие операторы поиска можно использовать в НКРЯ? Каков их функционал?
- В чем особенности корпусов М.Дэвиса?
- В чем особенности Ngram Viewer?
- В чем особенности Sketch Engine?
- Для чего используется AntConc? Каков его функционал?
- Каков функционал корпусного менеджера? Приведите примеры корпусных менеджеров.
- Каковы ограничения современных корпусных менеджеров?
- Как создать корпус самому? Каковы этапы? Для чего это может понадобиться? При помощи каких программ это можно сделать?
- Какие существуют статистические метрики для определения коллокаций? (t-score, MI-3, loglikelihood, logDice, etc.) Каковы сильные и слабые стороны каждой?
- Для чего применяется ipm? Как она рассчитывается?
- Что может являться единицей корпуса?
- Как отбираются тексты для корпуса? Проиллюстрируйте принципы отбора на примере любого корпуса.
- Представьте классификацию корпусов. Поясните, что означает «исследовательский корпус», «статический корпус», «параллельный корпус».
- Что такое конкордансер?
- Что такое дистрибуция?
- Что такое лексическая плотность?
- Что такое морфологический анализатор?
- Что такое сегментация корпуса?
- Назовите основные сферы применения корпусов.
- Что означает понятие накопительная сбалансированность?
- Чем отличается лингвистический корпус от электронного словаря?
- На каких ресурсах можно получить информацию о частотности употребления слова в конкретный период времени?
- В чем отличие corpus-based и corpus-driven исследований?
- Как можно оценить лексическое разнообразие?
- Какие новые корпуса (например, Common Crawl, mC4) используются для обучения LLM?
- Как LLM используются для автоматической аннотации и разметки корпусов?

Примеры практикоориентированных заданий:

- 1) провести мини-исследование по интересующей теме, связанной с возможностями применения инструментов компьютерной лингвистики
- 2) сравнить системы машинного перевода

- 3) исследовать предложенные тексты количественными методами
- 4) визуализировать предложенные тексты изученными методами.
- 5) построить граф по заданным критериям при помощи программы сетевого анализа Gephi.
- 6) провести стилеметрическое исследование.
- 7) найти в НКРЯ слова, изменившие смысл
- 8) проанализировать популярность хэштегов
- 9) собрать digital born данные с использованием инструмента Web scraper. Каждый студент получает задание на создание парсера для интересующего его сайта. Перед созданием парсера следует изучить возможности Web scraper, понять пайплайн работы, разобраться в структуре интересующего сайта.
- 10) провести тематическое моделирование отзывов игроков, проанализировать их, и попытаться составить портрет игрового опыта, связав это с особенностями разных игр
- 11) защита проекта исследования или ресурса по компьютерной лингвистике. В процессе защиты студент должен показать глубокое знакомство с теми из основных современных методов компьютерной лингвистики, которые применяются в её/его проекте.

Требования к выполнению заданий (или шкалы и критерии оценивания):

Для оценивания результатов обучения на зачете используется 4-балльная шкала: «отлично», «хорошо», «удовлетворительно» и «неудовлетворительно».

Соотношение показателей, критериев и шкалы оценивания результатов обучения.

Критерии оценивания компетенций	Уровень сформированности компетенций	Шкала оценок
<i>Обучающийся в полной мере демонстрирует</i> знание информационно-лингвистических технологий, технологий автоматической обработки естественного языка и искусственного интеллекта; принципов работы лингвистически ориентированных программных продуктов; основных типов задач обработки и анализа естественно-языковых текстов; основных видов автоматизированных систем обработки и анализа естественно-языковых текстов умение использовать лингвистически ориентированные программные системы; подбирать информационно-коммуникационные технологии для наиболее эффективного решения профессиональных задач; применять информационно-лингвистические технологии, технологии автоматической обработки естественного языка и искусственного интеллекта в соответствии с решаемой профессиональной задачей; проводить оценку качества и осуществлять выбор автоматизированной технологии обработки текстов в конкретных условиях решения прикладных	<i>Повышенный уровень</i>	<i>Отлично</i>

информационно-аналитических задач; применять автоматизированные технологии обработки текстов при решении прикладных информационно-аналитических задач, владение навыками работы с программными системами, реализующими автоматизированные технологии автоматической обработки естественного языка Обучающимся получено 80-100 баллов по итогам текущей успеваемости.		
Обучающийся имеет знание информационно-лингвистических технологий, технологий автоматической обработки естественного языка и искусственного интеллекта; принципов работы лингвистически ориентированных программных продуктов; основных типов задач обработки и анализа естественно-языковых текстов; основных видов автоматизированных систем обработки и анализа естественно-языковых текстов умение использовать лингвистически-ориентированные программные системы; подбирать информационно-коммуникационные технологии для наиболее эффективного решения профессиональных задач; применять информационно-лингвистические технологии, технологии автоматической обработки естественного языка и искусственного интеллекта в соответствии с решаемой профессиональной задачей; проводить оценку качества и осуществлять выбор автоматизированной технологии обработки текстов в конкретных условиях решения прикладных информационно-аналитических задач; применять автоматизированные технологии обработки текстов при решении прикладных информационно-аналитических задач, владение навыками работы с программными системами, реализующими автоматизированные технологии автоматической обработки естественного языка Обучающимся получено не менее 70 баллов по итогам текущей успеваемости.	Базовый уровень	Хорошо
Обучающийся демонстрирует частичное знание информационно-лингвистических технологий, технологий автоматической обработки естественного	Пороговый уровень	Удовлетворительно

языка и искусственного интеллекта; принципов работы лингвистически ориентированных программных продуктов; основных типов задач обработки и анализа естественно-языковых текстов; основных видов автоматизированных систем обработки и анализа естественно-языковых текстов умение использовать лингвистически-ориентированные программные системы; подбирать информационно-коммуникационные технологии для наиболее эффективного решения профессиональных задач; применять информационно-лингвистические технологии, технологии автоматической обработки естественного языка и искусственного интеллекта в соответствии с решаемой профессиональной задачей; проводить оценку качества и осуществлять выбор автоматизированной технологии обработки текстов в конкретных условиях решения прикладных информационно-аналитических задач; применять автоматизированные технологии обработки текстов при решении прикладных информационно-аналитических задач, владение навыками работы с программными системами, реализующими автоматизированные технологии автоматической обработки естественного языка Обучающимся получено не менее 60 баллов по итогам текущей успеваемости.		
<i>Ответ на контрольно-измерительный материал не соответствует любым четырем из перечисленных показателей. Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки.</i> Обучающимся получено менее 60 баллов по итогам текущей успеваемости.	–	Не удовлетворительно

20.2 Промежуточная аттестация

Промежуточная аттестация по дисциплине осуществляется с помощью следующих оценочных средств: *Собеседование по билетам к зачету, Практико-ориентированные задания*

Перечень вопросов к зачету

1. Автоматический анализ естественного языка. Общие проблемы автоматической обработки естественного языка.
2. Автоматическая обработка письменного текста. Автоматическое распознавание текста.

3. Текст как объект интеллектуального анализа. Классификация и кластеризация текстов.
4. Самообучающиеся системы: нейронные сети. Машинное обучение в лингвистике.
5. Визуализация данных. Современные автоматизированные средства визуализации данных.
6. Информационный поиск. Основные понятия информационного поиска. Лингвистические аспекты информационного поиска.
7. Современные информационно-поисковые системы: принципы организации. Фильтрация и ранжирование.
8. Применение статистических методов в лингвистических исследованиях. Частотность.
9. Электронные лингвистические ресурсы. Лингвистические базы данных. Построение и применение лингвистических ресурсов.
10. Цифровые библиотеки. Информационные системы в филологических задачах.
11. Открытые данные и проблема их обработки. Верификация данных. Большие данные в филологии и языкознании.
12. Корпусная лингвистика. Принципы организации корпуса.
13. Основные свойства корпуса. Классификации корпусов.
14. Современные корпуса текстов. Интернет как корпус. Национальный корпус русского языка и другие русскоязычные корпуса. Корпусы иных языков.
15. Обработка корпусных данных. Работа с параллельными корпусами. Структура и назначение параллельных корпусов в решении задач ИИ.
16. Компьютерная лексикография. Прикладные аспекты лексикографии.
17. Электронные словари: основные типы словарей и принципы их организации. Структура словарной статьи электронного словаря. Электронные словари, доступные в сети.
18. Частотные словари. Иноязычные словари. Лексическая информация в системах ИИ.
19. Автоматический перевод. История создания систем автоматического перевода. Современное состояние отрасли.
20. Машинный перевод. Машинный vs автоматизированный перевод.
21. Современные системы машинного перевода. Перспективы развития машинного перевода.
22. Методы оценки качества перевода. Сравнение систем машинного перевода. Программные средства помощи переводчику.
23. Методика контент-анализа. Проведение анкетирования.
24. Мультимодальная лингвистика. Работа с видео- и аудиоматериалами.
25. Экспертные системы. Стилеметрия в переводе.

Примеры практикоориентированных заданий:

- 1) провести мини-исследование по интересующей теме, связанной с возможностями применения инструментов компьютерной лингвистики
- 2) сравнить системы машинного перевода
- 3) исследовать предложенные тексты количественными методами
- 4) визуализировать предложенные тексты изученными методами.
- 5) построить граф по заданным критериям при помощи программы сетевого анализа Gephi.
- 6) провести стилеметрическое исследование.
- 7) найти в НКРЯ слова, изменившие смысл
- 8) проанализировать популярность хэштегов

- 9) собрать digital born данные с использованием инструмента Web scraper. Каждый студент получает задание на создание парсера для интересующего его сайта. Перед созданием парсера следует изучить возможности Web scraper, понять пайплайн работы, разобраться в структуре интересующего сайта.
- 10) провести тематическое моделирование отзывов игроков, проанализировать их, и попытаться составить портрет игрового опыта, связав это с особенностями разных игр
- 11) защита проекта исследования или ресурса по компьютерной лингвистике. В процессе защиты студент должен показать глубокое знакомство с теми из основных современных методов компьютерной лингвистики, которые применяются в её/его проекте.

Оценка знаний, умений и навыков, характеризующая этапы формирования компетенций в рамках изучения дисциплины осуществляется в ходе текущей и промежуточной аттестаций.

Текущая аттестация проводится в соответствии с Положением о текущей аттестации обучающихся по программам высшего образования Воронежского государственного университета. Текущая аттестация проводится в формах устного опроса (индивидуальный опрос, фронтальная беседа, доклады); тестирования, выполнения индивидуальных заданий. Критерии оценивания приведены выше.

Промежуточная аттестация проводится в соответствии с Положением о промежуточной аттестации обучающихся по программам высшего образования.

Контрольно-измерительные материалы промежуточной аттестации включают в себя теоретические вопросы, позволяющие оценить уровень полученных знаний, а также индивидуальные практические задания. При оценивании используются качественные шкалы оценок. Критерии оценивания приведены выше.

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

Перечень заданий для оценки сформированности компетенции:

- 1) закрытые задания (тестовые, средний уровень сложности):
 1. Что из перечисленного является корпусом?
 - a. Topic Modelling Tool
 - b. Google forms
 - c. GeoJSON
 - d. НКРЯ
 2. Какую программу Вы будете использовать для сетевого анализа?
 - a. Gephi
 - b. Linguae
 - c. YouAlign
 - d. Data Wrapper
 3. Какой ресурс Вы будете использовать для проведения стилеметрии?
 - a. R
 - b. SketchEngine
 - c. N-gram Viewer
 - d. Gephi
 4. Какую программу Вы будете использовать для анализа речи?
 - a. OPUS
 - b. Praat

- c. Hunalign
 - d. Voyant tools
5. Какую программу Вы будете использовать для распознавания текста (OCR)?
- a. PROMT
 - b. ABBYY Fine Reader Online
 - c. AntConc
 - d. ELAN
6. Какую программу Вы будете использовать для анализа собственного корпуса?
- a. Elibrary
 - b. GitHub
 - c. AntConc
 - d. Sublime
7. Какую программу Вы будете использовать для построения облаков слов?
- a. Skuuper Cleaner
 - b. Praat
 - c. Voyant tools
 - d. Zotero
8. Каким ресурсом Вы воспользуетесь, чтобы сравнить частоту использования существительного *a book* и глагола *to book* в английском языке с 1900 по 2010 гг.?
- a. Gephi
 - b. N-gram Viewer
 - c. Convertio
 - d. Hunalign
9. Метод, создающий базу для последующего морфологического и синтаксического анализа на основе выделения слов, цифровых комплексов, формул и т. д.
- a. Семантический анализ
 - b. Классификация
 - c. Графематический анализ
 - d. Кластеризация
10. Какая технология компьютерной лингвистики поможет сосчитать в тексте количество глаголов и прилагательных?
- a. морфологический анализатор
 - b. распознавание символов
 - c. анализ тональности
 - d. извлечение ключевых слов
11. Какая технология компьютерной лингвистики поможет сократить текст, убрав из него все причастные и деепричастные обороты?
- a. синтаксический анализ
 - b. извлечение именованных сущностей
 - c. анализ жанров
 - d. поиск плагиата
12. Какая технология компьютерной лингвистики поможет проанализировать комментарии к одной из нашумевших новостей и оценить отношение к событию?
- a. морфологический анализ
 - b. синтаксический анализ
 - c. реферирование текстов
 - d. анализ тональности
13. Какая технология компьютерной лингвистики поможет проанализировать массив текстов и понять, о чем они?

- a. поиск плагиата
 - b. извлечение именованных сущностей
 - c. распознавание символов
 - d. извлечение ключевых слов
14. Библиотека для получения векторных представлений слов на основе их совместной встречаемости в текстах.
- a. TEI
 - b. word2vec
 - c. tf-idf
 - d. regex
15. Модель, часто используемая при обработке текстов, представляющая собой неупорядоченный набор слов, входящих в обрабатываемый текст.
- a. word2vec
 - b. tf-idf
 - c. doc2vec
 - d. Bag of Words
16. Какое онлайн-приложение вы выберете для визуализации текста в виде облаков слов?
- a. AntConc
 - b. ELAN
 - c. Wordle
 - d. Praat
17. Автоматическое нахождение основы слова:
- a. Лемматизация
 - b. Стемминг
 - c. Кластеризация
 - d. Анализ тональности
18. Автоматическое приведение словоформ к начальной форме:
- a. Стемминг
 - b. Разметка
 - c. Лемматизация
 - d. Tf-idf
19. Метрика в задачах машинного обучения, являющаяся средним гармоническим между полнотой и точностью.
- a. F-Score
 - b. Precision
 - c. Recall
 - d. Accuracy
20. Какую программу Вы будете использовать для парсинга:
- a. SQL
 - b. AntConc
 - c. ParseHub
 - d. Topic Modelling Tool
21. Преимущества языковых корпусов – это:
- a. возможность хранения неограниченного объема текстовых данных
 - b. возможность многократного использования информации
 - c. возможность быстрого поиска и сбора данных
 - d. все перечисленное выше

22. Автоматический анализ структуры текстовых данных, который позволяет осуществлять процедуру сбора синтаксических конструкций (предложений) на естественном языке из Интернет-источников – это:
- Парсинг
 - Коллокация
 - Онтология
 - Тезаурус
23. Сведения, которые не являются частью текста, но содержат информацию о нем:
- Метаданные
 - Коллокации
 - Стоп-слова
 - N-граммы
24. Какую программу Вы будете использовать для выравнивания параллельных текстов?
- ABBY Aligner
 - Online OCR
 - WIX
 - Mendeley
25. Какой ресурс вы будете использовать для проведения тематического моделирования?
- YouAlign
 - OPUS
 - Topic Modelling Tool
 - Linguee
26. Семантический словарь, т. е. словарь, в котором представлены смысловые связи слов — синонимические, отношения Род-Вид (иногда называемые отношением Выше-Ниже), Часть-Целое, ассоциации, пр.
- База данных
 - Тезаурус
 - Корпус
 - Онтология
27. Набор понятий, сущностей определенной области знаний, ориентированный на многократное использование для различных задач.
- Онтология
 - Корпус
 - Лингвистическая разметка
 - Словарь
28. Что из перечисленного является примером онтологии?
- GloWbE
 - PyТез
 - НКРЯ
 - WordNet

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

- _____ системы – это программные системы для хранения, поиска и выдачи интересующей пользователя информации.

2. Синтаксический _____ – процедура приписывания грамматических характеристик цепочке слов.
3. _____ – это теория и практика разработки словарей.
4. Самостоятельное направление в прикладной лингвистике, ориентированное на использование компьютеров для решения задач, связанных с использованием естественного языка – это _____ лингвистика.
5. _____ поиска – отношение между количеством выданных релевантных текстов к общему количеству выданных системой текстов.
6. _____ поиска – соотношение между количеством выданных релевантных текстов или документов к общему количеству релевантных документов, имеющихся в данной информационной системе.
7. _____ лингвистика – это междисциплинарное прикладное направление, в котором объектом изучения является язык или речь, а инструментом анализа – количественные или статистические методы.
8. _____ – компьютерная коллекция текстов, специально подобранная и подготовленная для научных исследований.
9. История машинного перевода начинается с «_____ эксперимента» в январе 1954 г.
10. _____-слова – слова, знаки и символы, которые не обладают значением сами по себе, например, предлоги, частицы и местоимения.
11. _____ выражения – формальный язык поиска, основанный на использовании специальных метасимволов.
12. N-_____ – последовательности из N слов
13. _____ – этап, в рамках которого происходит разделение текста на более мелкие единицы — на предложения и слова.
14. Тематическое _____ – построение модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов.
15. _____ – это список всех употреблений заданного языкового выражения в контексте.
16. Text _____ – извлечения информации из неструктурированных текстовых данных.
17. Корпусный _____ – поисковая система по текстовым данным, которая предоставляет статистику о языковых единицах и приводит информацию в удобный для анализа вид.
18. Закон _____ гласит следующее: Если все слова языка (или просто относительно длинного текста) упорядочить по убыванию частоты их использования, то частота n-го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру n (так называемому рангу этого слова).
19. _____ — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов.
20. Мера _____ – метрика, показывающая все ли правильные примеры нашел классификатор и определяющаяся как количество найденных правильных примеров к общему количеству релевантных примеров в корпусе.
21. Мера _____ – метрика, отражающая количество релевантных категории примеров к общему количеству классифицированных как правильные примеров.
22. _____ текста – определение подлинности или подложности рукописного текста и установление его автора.

23. _____ – исследование и измерение стиливых характеристик текста с целью установления авторства или получения каких-либо сведений об авторе и условиях создания текстового документа.
24. Natural Language _____ – обработка естественного языка с помощью компьютерных технологий.
25. _____ (или аннотация) – это приписывание грамматической информации о входящих в тексты словоформам.
26. _____ текст (битекст) – текст на одном языке вместе с его переводом на другой язык.
27. _____ битекста – это сопоставление соответствующих друг другу предложений одного языка с синтаксическими единицами другого.

Критерии и шкалы оценивания заданий ФОС:

Для оценивания выполнения заданий используется балльная шкала:

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

- 1 балл – указан верный ответ;
- 0 баллов – указан неверный ответ (полностью или частично неверный).

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

- 2 балла – указан верный ответ;
- 0 баллов – указан неверный ответ (полностью или частично неверный).

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