

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное
учреждение высшего образования
«Государственный институт русского языка им. А.С. Пушкина»
Филологический факультет
Кафедра общего и русского языкознания

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ
ИНСТРУМЕНТЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ АНАЛИЗА
ЯЗЫКА И ТЕКСТА

Направление подготовки: 45.04.02 Лингвистика

Направленность (профиль): Лингвистика в условиях социокультурного
многообразия

Квалификация выпускника: магистр

Форма обучения: очная

Протокол заседания кафедры от «5» июля 2024 г. протокол № 16.

Протокол ученого совета филологического факультета от «30» августа 2024 г. протокол № 1.

Москва, 2024 год

Разработчик (и):

Р.Е. Тельпов, кандидат филологических наук, доцент кафедры общего и русского языкознания ФГБОУ ВО «Государственный институт русского языка им. А.С. Пушкина»

1. Цель освоения дисциплины

Целями освоения дисциплины является систематизация и расширение знаний в области новых информационных технологий, формирование у студента навыков использования методов статистической и автоматической обработки лингвистических явлений для решения прикладных задач в науке и осуществления грамотных лингвистических исследований.

2. Место дисциплины в структуре ОПОП

Дисциплина «Инструменты искусственного интеллекта для анализа языка и текста» входит в состав обязательной части основной образовательной программы магистратуры (Б1.О.13) и изучается в 1 семестре.

3. Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения ОПОП

В процессе изучения дисциплины формируются следующие компетенции:

Требования к результатам освоения дисциплины

Код и наименование компетенции	Код и наименование индикатора достижения компетенции (ИДК)	Результаты обучения по дисциплине		
		знать	уметь	владеть
ОПК-3 Способен применять в профессиональной деятельности общедидактические принципы обучения и воспитания, использовать современные методики и технологии организации образовательного процесса;	ОПК-3.2. Адекватно определяет методический потенциал современных электронных обучающих платформ (систем управления обучением) для разработки электронных учебных материалов по иностранным языкам	- основные понятия количественной лингвистики; - основные базы данных и информационно-поисковые системы, а также методы работы с этими системами и базами данных	- работать с основными информационно-поисковыми и экспертными системами, системами представления знаний, синтаксического и морфологического анализа, автоматического синтеза, распознавания и понимания речи	- навыками обработки лексикографической информации и автоматизированного перевода, автоматизированными системами идентификации и верификации личности
ОПК-7 Способен работать с основными информационно-поисковыми и экспертными системами, системами представления знаний и обработки вербальной информации	ОПК-7.1. Корректно использует профильные информационные ресурсы информационно-телекоммуникационной сети «Интернет»	о системных потенциях языка с опорой на количественные методы исследования языковых единиц	решать задачи определения авторского инварианта	навыками контент-анализа с использованием частотных характеристик языковых единиц
	ОПК-7.2. Владеет рациональными приемами поиска и	принципы создания и функционирования	пользоваться современным лексикографическим	навыками использования современного лексикографическим

	применения программных продуктов лингвистического профиля	лингвистических компьютерных программ	инструментарием	ого инструментария (электронные словари, сетевые словари, тезаурусы) для оптимизации лингвистических исследований
--	---	---------------------------------------	-----------------	---

4. Объем, структура и содержание дисциплины

4.1. Трудоемкость дисциплины

Вид учебной работы	очная форма обучения
	1 семестр
Общая трудоемкость по учебному плану (з.е.)	3
Количество часов по учебному плану (час.)	108
Контактная работа (всего)	24
В том числе:	
Лекции (Л)	8
Практические занятия (ПЗ)	16
Лабораторные работы (ЛР)	
Самостоятельная работа (всего)	82
Вид промежуточной аттестации - зачет	2

4.2. Структура и содержание дисциплины

Наименование и краткое содержание разделов дисциплины	Количество часов по учебному плану	Контактная работа (часы), из них		Самостоятельная работа	Форма промежуточной аттестации	Формы текущего контроля успеваемости
		Лекции	Практические занятия			
	Очная форма	Очная форма	Очная форма	Очная форма	Очная форма	Очная форма
1. Методы статистической обработки данных в лингвистических исследованиях Расчет стандартизированных остатков. Список Сводеша как объект статистического исследования. Закон Ципфа. Частотные словари.	30	4	4	22		УО, П, Д
2. Корпусная лингвистика. Национальные корпуса текстов. Корпусная лингвистика как наука. Понятие «корпус текста». Классификация корпусов. Особенности применения корпусов текстов. Национальные корпуса текстов	36	2	4	30		УО, П, Д
3. Чат-бот VS нейросеть. Особенности речи, сгенерированной искусственным интеллектом История сознания чат-ботов и нейросетей. Чем чат-бот отличается от нейросети. Особенности речи, сгенерированной искусственным интеллектом (синтаксический, морфологический, лексический, фонетический и статистический аспекты). Об одушевленности слова «чат-бот». Методы распознавания и обработки текстов: токенизация, «мешок слов», регулярные выражения, автоматическая разметка семантических ролей. Тест Тьюринга. Элиза и Пэрри – первые чат боты. Методы тестирования и моделирования речемыслительных способностей искусственного интеллекта (метафора «португальского стадиона», метафора «китайской комнаты»).	40	2	8	30		УО, П, Д
Форма промежуточной аттестации – <u>зачет</u>	2				2	
Итого	108	8	16	82	2	

Примечание:

Условные обозначения форм контроля: УО – устный опрос; П – проект; Д – доклад

5. Образовательные технологии

В преподавании дисциплины «Инструменты искусственного интеллекта и его значение для анализа языка и текста» используются разнообразные образовательные технологии, в том числе:

- ~ лекции;
- ~ практические занятия;
- ~ выступления с докладами.

Самостоятельная работа охватывает проработку обучающимися отдельных вопросов теоретического курса, выполнение домашних работ, включающих

- ~ подготовку презентаций, докладов;
- ~ работа с учебной и научной литературой.

6. Учебно-методическое обеспечение самостоятельной работы обучающихся

6.1. Список учебной литературы

Основная литература:

1. Воронов, М. В. Системы искусственного интеллекта: Учебник и практикум для вузов / М. В. Воронов, В. И. Пименов, И. А. Небаев. — 2-е изд., перераб. и доп. — Москва: Издательство Юрайт, 2024. — 268 с. — (Высшее образование). — ISBN 978-5-534-17032-0. — Текст: электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://urait.ru/bcode/544161> (дата обращения: 06.12.2024).

Дополнительная литература:

1. Станкевич, Л. А. Интеллектуальные системы и технологии: Учебник и практикум для вузов / Л. А. Станкевич. — 2-е изд., перераб. и доп. — Москва: Издательство Юрайт, 2024. — 478 с. — (Высшее образование). — ISBN 978-5-534-20363-9. — Текст: электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://urait.ru/bcode/558009> (дата обращения: 06.12.2024).

6.2. Словари и справочники:

1. Большая Российская энциклопедия. - <https://old.bigenc.ru/> (Электронная версия издания с открытым доступом в Интернет)

6.3. Периодические издания:

-

6.4. Ресурсы информационно-телекоммуникационной сети «Интернет»:

Ресурсы сети Интернет, необходимые для изучения дисциплины:

Наименование	Ссылка на ресурс	Доступность (свободный доступ / ограниченный доступ)
Лингвистические корпуса и электронные собрания текстов		
Национальный корпус русского языка	www.ruscorpora.ru	Свободный
Электронная библиотека ИРЛИ РАН	http://lib.pushkinskijdom.ru/?ysclid=lyctg57lca218433412	Свободный
The Corpus of Contemporary American English (COCA)	www.english-corpora.org	Свободный
WordProject	https://www.wordproject.org/	Свободный
Электронные словари		
О.Н. Ляшевская, С.А. Шаров Новый частотный словарь русской лексики	http://dict.ruslang.ru/freq.php	Свободный
Словари русского языка – Грамота.ру	https://gramota.ru/biblioteka/slovari	Свободный
Электронно-библиотечные системы		
ЭБС Юрайт	https://urait.ru	Требуется регистрация
Ресурсы для анализа и обработки текстов		

Ваал-мини	http://vaal.ru/prog/free.php	Свободный
Лингвоанализатор	https://www.rusf.ru	Свободный
Онлайн-калькулятор	https://math.semestr.ru/index.php	Свободный
Текстометр	https://textometr.ru	Свободный
AntConc	https://laurenceanthony.net/software/antconc	Свободный
Gerwin AI — Сервис генеративного контента / Нейросеть	https://gerwin.io	Свободный
GIGA CHAT	https://developers.sber.ru	Свободный
SEO-анализ текста	https://pr-cy.ru/tools/	Свободный
WebScript.Ru	https://webscript.ru	Свободный

6.5. Содержание самостоятельной работы

<u>Методы статистической обработки данных в лингвистических исследованиях</u>
<u>Вопросы для самостоятельной работы:</u> 1. Критерий хи-квадрат. 2. Расчет среднеквадратичного отклонения. 3. Расчет стандартизированных остатков. 4. Список Сводеша как объект статистического исследования. 5. Закон Ципфа: история открытия, вариации, формулировка, применение. 6. Частотные словари. Понятие об IPM. Понятие о «коэффициенте Жуайена». «Метафора хоббита». <u>Задания для самостоятельной работы:</u> 1. Проведите статистический анализ тем какого-либо периодического издания с применением «критерия хи-квадрат», с использованием метода стандартизированных остатков и с использованием метода расчета среднеквадратичного отклонения (образец см. в ФОМ) 2. Опираясь на закон Ципфа проанализируйте лексический состав какого-либо цельного текста, выявите его статистические характеристики. О чем они свидетельствуют? 3. Пользуясь частотным словарем проанализируйте динамику в изменениях нескольких любых слов, относящихся к разным пластам лексики (абстрактные, конкретные, историзмы, архаизмы и т.д.).
<u>Корпусная лингвистика. Национальные корпуса текстов</u>
<u>Вопросы для самостоятельной работы:</u> 1. Корпусная лингвистика. 2. Национальные корпуса текстов. 3. Корпусная лингвистика как наука. 4. Понятие «корпус текста». 5. Классификация корпусов. 6. Особенности применения корпусов текстов. 7. Национальные корпуса текстов <u>Задания для самостоятельной работы:</u> 1. Пользуясь Национальным корпусом русского языка проанализируйте трансформацию в значении какого-либо слова в диахронном аспекте. Напишите на эту тему исследование (образец см. в ФОМ). 2. На примере нескольких слов, объединенных системными отношениями (синонимов, омонимов, антонимов), сравните сходства и различия двух подкорпусов Национального корпуса русского языка. 3. На примере нескольких слов, объединенных системными отношениями (синонимов, омонимов, антонимов), сравните сходства и различия двух лингвистических корпусов – НКРЯ и СОСА, сравните их возможности.
<u>Чат-бот VS нейросеть. Особенности речи, сгенерированной искусственным интеллектом</u>

Вопросы для самостоятельной работы

1. История сознания чат-ботов и нейросетей. Чем чат-бот отличается от нейросети. Элиза и Пэрри – первые чат боты.
2. Особенности речи, сгенерированной искусственным интеллектом (синтаксический, морфологический, лексический, фонетический и статистический аспекты).
3. Грамматические характеристики слова «чат-бот».
4. Методы распознавания и обработки текстов: токенизация, «мешок слов», регулярные выражения, автоматическая разметка семантических ролей.
5. Тест Тьюринга. Методы тестирования и моделирования речемышлительных способностей искусственного интеллекта (метафора «португальского стадиона», метафора «китайской комнаты»).

Задания для самостоятельной работы

1. Сравните две группы текстов – созданных человеком и созданных нейросетью. Попробуйте выявить отличия между ними с опорой на их морфологические, лексические, графические и синтаксические особенности.
2. Сравните две группы текстов – созданных человеком и созданных нейросетью. Попробуйте выявить отличия между ними с опорой на их статистические характеристики.
3. Сравните тексты, созданные разными нейросетями на основе одних и тех же промптов.

7. Оценочные средства для проведения текущего контроля успеваемости и промежуточной аттестации обучающихся

Оценка качества учебных достижений обучающихся по дисциплине осуществляется в виде текущего контроля успеваемости и промежуточной аттестации.

7.1. Текущий контроль успеваемости

Текущий контроль успеваемости учитывает следующие виды текущей аттестации: выполнение обучающимся всех видов контрольных мероприятий, предусмотренных рабочей программой дисциплины, и самостоятельную работу обучающегося. Все формы текущего контроля и активность участия студентов в практических и (или) лабораторных занятиях, критерии их оценивания представлены в рабочей программе дисциплины, технологической карте и фонде оценочных материалов (Приложения 1-2).

7.1.1. Тематика и требования к структуре и содержанию докладов

Примерная тематика докладов:

1. Основная проблематика и методы прикладной лингвистики. Задачи и сферы применения квантитативной лингвистики.
2. Значимость статистической обработки данных для решения исследовательских и практических задач.
3. Дешифровка сообщений/текстов и сферы ее применения. Актуальность атрибуции текстов в современных условиях.
4. Электронные словари и основания для их классификации. Задачи и сферы применения корпусной лингвистики.
5. Автоматический анализ текста и области его применения.
6. Автоматизированные системы перевода. Машинный фонд языка.
7. Проблемы формирования базы данных на основе электронных словарей. Проблемы формирования базы данных на основе корпуса национального языка.
8. Возможности использования квантитативной лингвистики и новых информационных технологий для решения проблем межкультурной коммуникации.
9. Возможности использования квантитативной лингвистики и новых информационных технологий для решения проблем перевода.
10. Возможности использования квантитативной лингвистики и новых информационных технологий для решения проблем обучения языкам.
11. Кибернетика и лингвистика.
12. Формальные языки и естественные языки: сходства и отличия.
13. Основные методы и понятия стилистики и атрибуции текста.
14. Извлечение морфологической информации по Национальному корпусу русского языка.

Требования к структуре и содержанию, критерии и шкала оценки доклада в баллах:

Критерии	Показатели
Выполнены все требования к написанию доклада: тема раскрыта полностью, сформулированы выводы, выдержан объем, соблюдены требования к внешнему оформлению, даны правильные ответы на	15 баллов

дополнительные вопросы	
Основные требования к докладу выполнены, но при этом имеются недочеты. В частности, имеются неточности в изложении материала, отсутствует логическая последовательность в суждениях, не выдержан объем доклада, имеются упущения в оформлении, на дополнительные вопросы даны неполные ответы	10 баллов
Имеются существенные отступления от требований к оформлению доклада: тема освещена лишь частично, допущены фактические ошибки в содержании или при ответе на дополнительные вопросы (или не смог ответить на дополнительные вопросы), отсутствует вывод	5 баллов
Обучающийся не смог раскрыть тему доклада, обнаруживается существенное непонимание проблемы, не ответил на дополнительные вопросы либо доклад не представлен	1-0 баллов

7.1.2. Тематика и требования к структуре и содержанию эссе

Данный вид работ не предусмотрен.

7.1.3. Тематика и требования к структуре и содержанию рефератов

Данный вид работ не предусмотрен.

7.1.4. Тематика и требования к структуре и содержанию контрольных работ

Данный вид работ не предусмотрен.

7.1.5. Тематика и требования к структуре и содержанию тестов

Данный вид работ не предусмотрен.

7.1.6. Тематика и требования к структуре и содержанию проектов:

Тематика проектов:

1. Проведите статистический анализ тем какого-либо периодического издания с применением «критерия хи-квадрат», с использованием метода стандартизированных остатков и с использованием метода расчета среднеквадратичного отклонения (образец см. в ФОМ).

2. Опираясь на закон Ципфа проанализируйте лексический состав какого-либо цельного текста, выявите его статистические характеристики. О чем они свидетельствуют?

3. Пользуясь частотным словарем проанализируйте динамику в изменениях нескольких любых слов, относящихся к разным пластам лексики (абстрактные, конкретные, историзмы, архаизмы и т.д.).

4. Пользуясь Национальным корпусом русского языка проанализируйте трансформацию в значении какого-либо слова. Напишите на эту тему исследование (образец см. в ФОМ).

5. На примере нескольких слов, объединенных системными отношениями (синонимов, омонимов, антонимов), сравните сходства и различия двух лингвистических корпусов – НКРЯ и СОСА, сравните их возможности.

6. Сравните две группы текстов – созданных человеком и созданных нейросетью. Попробуйте выявить отличия между ними с опорой на их морфологические, лексические, графические и синтаксические особенности.

7. Сравните две группы текстов – созданных человеком и созданных нейросетью. Попробуйте выявить отличия между ними с опорой на их статистические характеристики.

8. Сравните тексты, созданные разными нейросетями на основе одних и тех же промптов.

Критерии и шкала оценки проекта в баллах:

Критерии	Показатели
Выполнены все требования к написанию проекта: тема раскрыта полностью, сформулированы выводы, выдержан объем, соблюдены требования к внешнему оформлению, даны правильные ответы на дополнительные вопросы	5 баллов
Основные требования к проекту выполнены, но при этом имеются недочеты. В частности, имеются неточности в изложении материала, отсутствует логическая последовательность в суждениях, не выдержан объем доклада, имеются упущения в оформлении, на дополнительные вопросы даны неполные ответы	4 балла
Имеются существенные отступления от требований к оформлению проекта: тема освещена лишь частично, допущены фактические ошибки в содержании или при ответе на дополнительные вопросы (или не смог ответить на дополнительные вопросы), отсутствует вывод	3 балла
Обучающийся не смог раскрыть тему проекта, обнаруживается существенное непонимание проблемы, не ответил на дополнительные вопросы либо проект не представлен	2 балла

7.2. Промежуточная аттестация

Промежуточная аттестация по дисциплине обеспечивает оценку степени полноты и качества освоения компетенций в соответствии с планируемыми результатами обучения по дисциплине.

7.2.1. Оценивание результатов освоения образовательной программы в части дисциплин (модулей), практик

Результаты освоения дисциплины оцениваются 1) по 4-балльной системе: отлично, хорошо, удовлетворительно, неудовлетворительно (экзамен, зачет с оценкой); 2) по шкале зачтено, не зачтено (зачет).

Текущий контроль, промежуточная аттестация осуществляются в соответствии с локальным нормативным актом Института.

Оценочные материалы для проведения промежуточной аттестации обучающихся приведены в Приложении 2 к рабочей программе дисциплины.

7.2.2. Примерные вопросы для подготовки к зачету:

1. Что изучает количественная лингвистика?

2. Критерий хи-квадрат? Зачем используется данный критерий в статистических исследованиях?

3. Фактор в статистических исследованиях? Уровень фактора в статистических исследованиях? Отклик в статистических исследованиях? Поясните эти термины на примере конкретного исследования, составив таблицу сопряженности.

4. Расчет среднеквадратичного отклонения в статистических исследованиях. Как расчет среднеквадратичного отклонения соотносится с критерием хи-квадрат? Что такое генеральная лингвистическая совокупность? Что такое выборочная лингвистическая совокупность?

5. Что дает для понимания статистических закономерностей метод расчета стандартизированных остатков?

6. Закон Ципфа. Почему именно Джорджа Ципфа называют отцом количественной лингвистики? «Метафора хоббита».

7. Атрибуция текста. В какой сфере она применяется? Какие виды атрибуции текста вы знаете? Чем приемы атрибуции древних рукописных памятников отличаются от атрибуции современных текстов? Какие вы знаете современные электронные ресурсы, помогающие установить авторство текста

8. Основные вехи в развитии стилистики как науки.

9. Лингвистический корпус: его предназначение, виды.

10. Основные вехи в истории корпусной лингвистики.

11. Направления автоматизированной обработки текста на естественных языках. Искусственный интеллект. Основные этапы развития идеи искусственного интеллекта.

12. Реферат и аннотация. Основные характерные черты и виды этих документов. Основные этапы реферирования и аннотирования текста.

7.2.3. Примерное задание к зачету:

Заданием к зачету служит либо работа над ошибками, допущенными в текущих заданиях, либо выполнение работ, не выполненных студентами ранее.

7.2.4.

Критерии оценивания результатов обучения по дисциплине при проведении промежуточной аттестации:

Планируемые результаты обучения по дисциплине	Критерии оценивания результатов обучения по дисциплине			
	«неудовлетворительно»	«удовлетворительно»	«хорошо»	«отлично»
	«не зачтено»	«зачтено»		
Сформированность знаний (знать)	Уровень знаний ниже минимальных требований. Имели место грубые ошибки. Невозможно оценить полноту знаний вследствие отказа обучающегося от ответа.	Минимально допустимый уровень знаний. Допущено много негрубых ошибок.	Уровень знаний в объеме, соответствующем программе подготовки. Допущено несколько несущественных ошибок.	Уровень знаний в объеме, соответствующем программе подготовки, или в объеме, превышающем программу подготовки, продемонстрирован без ошибок и погрешностей.
Сформированность умений (уметь)	Отсутствие минимальных умений. При решении стандартных задач не продемонстрированы основные умения. Имели место грубые ошибки. Невозможно оценить наличие умений вследствие отказа обучающегося от ответа.	Продemonстрированы основные умения. Решены типовые задачи с негрубыми ошибками. Выполнены все задания, но не в полном объеме.	Продemonстрированы все основные умения. Решены все основные задачи, но некоторые с негрубыми ошибками. Выполнены все задания, в полном объеме, но некоторые с недочетами.	Продemonстрированы все основные умения. Решены все основные задачи. Выполнены все задания, в полном объеме без недочетов.

Сформированность навыков (владеть)	При решении стандартных задач не продемонстрированы базовые навыки.	Продemonстрирован минимальный набор навыков.	Продemonстрированы базовые навыки.	Продemonстрировано всестороннее владение навыками.
	Имели место грубые ошибки. Невозможно оценить наличие навыков вследствие отказа обучающегося от ответа.	Решены стандартные задачи с некоторыми недочетами.	Решены стандартные задачи, но некоторые с недочетами.	Задачи, в том числе нестандартные, решены без ошибок и недочетов.

8. Материально-техническое обеспечение дисциплины

Для реализации рабочей программы дисциплины имеются помещения для проведения занятий лекционного и семинарского типа, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы. Помещения представляют собой учебные аудитории для проведения учебных занятий, оснащенные оборудованием и техническими средствами обучения, служащими для представления информации большой аудитории.

Помещение для проведения занятий лекционного типа поточная аудитория на 100 посадочных мест укомплектовано:

Помещение для проведения занятий семинарского типа, текущего контроля и промежуточной аттестации укомплектовано: специализированной мебелью, компьютером, интерактивной доской, мультимедиа проектором, звуковой системой.

Помещение для самостоятельной работы оснащено компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа к электронной информационно-образовательной среде Института.

Институт обеспечен следующим необходимым комплектом лицензионного и свободно распространяемого программного обеспечения, в том числе отечественного производства:

1. Microsoft Windows 10,
2. Microsoft Office 2016 стандартная,
3. Big Blue Button.

