

**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ**

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Воронежский государственный технический университет»

УТВЕРЖДАЮ

Декан факультета информационных
технологий и компьютерной безопасности

Гусев П.Ю.

«21» декабря 2021 г.



РАБОЧАЯ ПРОГРАММА

дисциплины

«Технологии работы с естественным языком Nature Language
Processing»

Направление подготовки 09.04.01 Информатика и вычислительная техника

Профиль Искусственный интеллект

Квалификация выпускника магистр

Нормативный период обучения 2 года / 2 года и 5 м.

Форма обучения очная / заочная

Год начала подготовки 2022

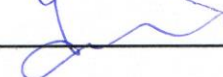
Автор программы


_____/Гусев П.Ю./

Заведующий кафедрой
Компьютерных
интеллектуальных
технологий проектирования


_____/М.И. Чижов/

Руководитель ОПОП


_____/М.И. Чижов/

Воронеж 2021

1. ЦЕЛИ И ЗАДАЧИ ДИСЦИПЛИНЫ

1.1. Цели дисциплины

изучение студентами с основными приложениями, подходами, источниками данных и инструментами обработки естественного языка; обучение навыкам решения задач, связанных с обработкой текстов, а также оценке таких решений.

Изучение дисциплины должно способствовать формированию у студентов основ научного мышления, в том числе: владение основными методами, способами и инструментами создания программного обеспечения, использования для решения практических задач

1.2. Задачи освоения дисциплины

- ознакомление с тенденцией развития программного обеспечения;
- изучение технологий работы с естественным языком;
- изучение средств реализации программного обеспечения;
- изучение методик и подходов к обработке естественного обеспечения;
- изучение сфер применения интеллектуальных подсистем обработки естественного языка.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП

Дисциплина «Технологии работы с естественным языком Nature Language Processing» относится к дисциплинам части, формируемой участниками образовательных отношений (дисциплина по выбору) блока Б1.

3. ПЕРЕЧЕНЬ ПЛАНИРУЕМЫХ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ

Процесс изучения дисциплины «Технологии работы с естественным языком Nature Language Processing» направлен на формирование следующих компетенций:

УК-1 - Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, выработать стратегию действий

УК-2 - Способен управлять проектом на всех этапах его жизненного цикла

ОПК-10 - Способен адаптировать и применять на практике классические и новые научные принципы и методы исследований для решения задач в области создания и применения технологий и систем искусственного интеллекта и методы исследований

ПК-2 - Способен участвовать в решении профессиональных проектных задач, выбирать и реализовывать командную роль в работе над проектом в соответствии с приоритетами собственной деятельности

Компетенция	Результаты обучения, характеризующие сформированность компетенции
УК-1	Знает принципы разработки стратегии достижения поставленной цели как последовательность шагов, предвидя результат каждого шага и оценивая их влияние на внешнее окружение планируемой

	<p>деятельности на взаимоотношения участников этой деятельности</p> <p>Знает как осуществляется поиск вариантов решения поставленной задачи на основе доступных источников информации; определяет в рамках выбранного алгоритма вопросы, подлежащие дальнейшей разработке</p> <p>Владеет навыками анализа проблемной ситуации как системы, выявляя ее составляющие и связи между ними</p>
УК-2	<p>Знать принципы формирования результатов проекта (или отдельных его этапов) в форме отчета, статьи, выступления на научнопрактической конференции</p> <p>Уметь организовать и координировать работу участников проекта, обеспечивает работу команды необходимыми ресурсами</p> <p>Владеть навыками формирования план-графика реализации проекта в целом и контролирует его выполнение</p>
ОПК-10	<p>Знать методологии разработки программного обеспечения, методологии управления проектами разработки программного обеспечения, методы и средства организации проектных данных, лучшие практики управления разработкой программного обеспечения, нормативно-технические документы (стандарты и регламенты), описывающие процессы управления инфраструктурой коллективной среды разработки</p> <p>Уметь применять методологии разработки программного обеспечения, методологии управления проектами разработки программного обеспечения, методы и средства организации проектных данных, лучшие практики и отражать их в базе знаний, основные принципы и методы управления персоналом, нормативно-технические документы (стандарты и регламенты), описывающие процессы управления инфраструктурой коллективной среды разработки</p> <p>Владеть навыками выбора инструментальных средств разработки, определения набора библиотек повторно используемых модулей, выбора средств создания и ведения репозитория, учета задач, сборки и непрерывной интеграции, базы знаний</p>
ПК-2	<p>Знать сквозные цифровые технологии искусственного интеллекта, включая «Компьютерное зрение», «Обработка естественного языка», «Рекомендательные</p>

	системы и системы поддержки принятия решений», «Распознавание и синтез речи»; современный опыт применения систем искусственного интеллекта
	Уметь проводить сравнительный анализ методов и инструментальных средств для решения задач искусственного интеллекта; разрабатывать и оценивать модели искусственного интеллекта; программировать на языках высокого уровня, ориентированных на применение методов и алгоритмов искусственного интеллекта.
	Владеть навыками выявления требований заказчика к результатам решения задач искусственного интеллекта; определения возможностей применения методов искусственного интеллекта в предметной области решаемой задачи; использования имеющейся методологической и технологической инфраструктуры анализа и обработки данных.

4. ОБЪЕМ ДИСЦИПЛИНЫ

Общая трудоемкость дисциплины «Технологии работы с естественным языком Nature Language Processing» составляет 3 з.е.

Распределение трудоемкости дисциплины по видам занятий
очная форма обучения

Виды учебной работы	Всего часов	Семестры
		3
Аудиторные занятия (всего)	36	36
В том числе:		
Лекции	16	16
Лабораторные работы (ЛР)	20	20
Самостоятельная работа	72	72
Виды промежуточной аттестации - зачет	+	+
Общая трудоемкость:	час	108
	зач.ед.	3
		108
		3

заочная форма обучения

Виды учебной работы	Всего часов	Семестры
		4
Аудиторные занятия (всего)	16	16
В том числе:		
Лекции	8	8
Лабораторные работы (ЛР)	8	8

Самостоятельная работа		88	88
Часы на контроль		4	4
Виды промежуточной аттестации - зачет		+	+
Общая трудоемкость:	час	108	108
	зач.ед.	3	3

5. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

5.1 Содержание разделов дисциплины и распределение трудоемкости по видам занятий

очная форма обучения

№ п/п	Наименование темы	Содержание раздела	Лекц	Лаб. зан.	СРС	Всего, час
1	Обработка естественного языка	Типы задач обработки естественного языка. Традиционные методы и методы глубинного обучения	2	-	12	20
2	Векторное представление.	Представления для слов естественного языка. Векторное представление. Целевая функция векторизации слов. Построение векторизованной модели словаря. Сокращение размерности словаря	2	4	12	20
3	Методы обучения.	Методы обучения. Задачи классификации. Учет структуры предложения. Безконтекстная грамматика. Проблема неопределенности модификатора. Типы зависимостей. Задача разбора зависимостей. Оценка качества. SyntaxNet	2	4	12	18
4	Рекуррентные нейронные сети	Статистическая модель языка. N-граммы. Рекуррентные нейронные сети. Проблемы обучения и подходы к их решению. Модификации рекуррентных сетей: LSTM и GRU. Задачи машинного перевода. Оценка качества перевода. Отображение последовательностей (seq2seq)	4	4	12	18
5	Свёрточные нейронные сети	Свёрточные нейронные сети. Применение в системах анализа естественного языка. Работа с частями слов. Буквенные (character-based) модели. Гибридные модели. Контекстные представления. Проблема неизвестных слов.	4	4	12	16
6	Механизмы внимания	Механизмы внимания, их реализация, применение. Самовнимание. Трансформеры. Перенесение обучения между языковыми моделями	2	4	12	16
Итого			16	20	72	108

заочная форма обучения

№ п/п	Наименование темы	Содержание раздела	Лекц	Лаб. зан.	СРС	Всего, час
1	Обработка естественного языка	Типы задач обработки естественного языка. Традиционные методы и методы глубинного обучения	2	-	14	18
2	Векторное представление.	Представления для слов естественного языка. Векторное представление. Целевая функция векторизации слов. Построение векторизованной модели словаря. Сокращение размерности словаря	2	1	14	18
3	Методы обучения.	Методы обучения. Задачи классификации. Учет структуры предложения. Безконтекстная грамматика. Проблема неопределенности модификатора. Типы зависимостей. Задача разбора зависимостей. Оценка качества. SyntaxNet	2	2	14	18

4	Рекуррентные нейронные сети	Статистическая модель языка. N-граммы. Рекуррентные нейронные сети. Проблемы обучения и подходы к их решению. Модификации рекуррентных сетей: LSTM и GRU. Задачи машинного перевода. Оценка качества перевода. Отображение последовательностей (seq2seq)	2	2	14	18
5	Свёрточные нейронные сети	Свёрточные нейронные сети. Применение в системах анализа естественного языка. Работа с частями слов. Буквенные (character-based) модели. Гибридные модели. Контекстные представления. Проблема неизвестных слов.	-	2	16	16
6	Механизмы внимания	Механизмы внимания, их реализация, применение. Самовнимание. Трансформеры. Перенесение обучения между языковыми моделями	-	1	16	16
Итого			8	8	88	104

5.2 Перечень лабораторных работ

1. Учет структуры предложения. Безконтекстная грамматика. Проблема неопределенности модификатора. Типы зависимостей. Задача разбора зависимостей. Оценка качества. SyntaxNet

2. Статистическая модель языка. N-граммы. Рекуррентные нейронные сети. Проблемы обучения и подходы к их решению

3. Задачи машинного перевода. Оценка качества перевода. Отображение последовательностей (seq2seq)

4. Свёрточные нейронные сети. Применение в системах анализа естественного языка.

5. Работа с частями слов. Буквенные (character-based) модели. Гибридные модели

6. ПРИМЕРНАЯ ТЕМАТИКА КУРСОВЫХ ПРОЕКТОВ (РАБОТ) И КОНТРОЛЬНЫХ РАБОТ

В соответствии с учебным планом освоение дисциплины не предусматривает выполнение курсового проекта (работы) или контрольной работы.

7. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ПРОВЕДЕНИЯ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ

7.1. Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания

7.1.1 Этап текущего контроля

Результаты текущего контроля знаний и межсессионной аттестации оцениваются по следующей системе:

«аттестован»;

«не аттестован».

Компетенция	Результаты обучения, характеризующие сформированность компетенции	Критерии оценивания	Аттестован	Не аттестован
УК-1	Знает принципы разработки	Количество выполненных лабораторных работ	Выполнение работ в срок,	Невыполнение работ в срок,

	стратегии достижения поставленной цели как последовательность шагов, предвидя результат каждого шага и оценивая их влияние на внешнее окружение планируемой деятельности на взаимоотношения участников этой деятельности		предусмотренный в рабочих программах	предусмотренный в рабочих программах
	Знает как осуществляется поиск вариантов решения поставленной задачи на основе доступных источников информации; определяет в рамках выбранного алгоритма вопросы, подлежащие дальнейшей разработке	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	Владет навыками анализа проблемной ситуации как системы, выявляя ее составляющие и связи между ними	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
УК-2	Знать принципы формирования результатов проекта (или отдельных его этапов) в форме отчета, статьи, выступления на научнопрактической конференции	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	Уметь организовать и координировать работу участников проекта, обеспечивает работу команды необходимыми ресурсами	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	Владеть навыками формирования плана-графика реализации проекта в целом и контролирует его выполнение	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
ОПК-10	Знать методологии разработки программного обеспечения, методологии управления проектами разработки программного обеспечения, методы и средства организации проектных данных, лучшие практики управления разработкой программного обеспечения, нормативно-технические документы (стандарты и регламенты), описывающие процессы	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах

	управления инфраструктурой коллективной среды разработки			
	Уметь применять методологии разработки программного обеспечения, методологии управления проектами разработки программного обеспечения, методы и средства организации проектных данных, лучшие практики и отражать их в базе знаний, основные принципы и методы управления персоналом, нормативно-технические документы (стандарты и регламенты), описывающие процессы управления инфраструктурой коллективной среды разработки	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	Владеть навыками выбора инструментальных средств разработки, определения набора библиотек повторно используемых модулей, выбора средств создания и ведения репозитория, учета задач, сборки и непрерывной интеграции, базы знаний	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
ПК-2	Знать сквозные цифровые технологии искусственного интеллекта, включая «Компьютерное зрение», «Обработка естественного языка», «Рекомендательные системы и системы поддержки принятия решений», «Распознавание и синтез речи»; современный опыт применения систем искусственного интеллекта	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	Уметь проводить сравнительный анализ методов и инструментальных средств для решения задач искусственного интеллекта; разрабатывать и оценивать модели искусственного интеллекта;	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах

	программировать на языках высокого уровня, ориентированных на применение методов и алгоритмов искусственного интеллекта.			
	Владеть навыками выявления требований заказчика к результатам решения задач искусственного интеллекта; определения возможностей применения методов искусственного интеллекта в предметной области решаемой задачи; использования имеющейся методологической и технологической инфраструктуры анализа и обработки данных.	Количество выполненных лабораторных работ	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах

7.1.2 Этап промежуточного контроля знаний

Результаты промежуточного контроля знаний оцениваются в 3 семестре для очной формы обучения, 4 семестре для заочной формы обучения по двухбалльной системе:

«зачтено»

«не зачтено»

Компетенция	Результаты обучения, характеризующие сформированность компетенции	Критерии оценивания	Зачтено	Не зачтено
УК-1	Знает принципы разработки стратегии достижения поставленной цели как последовательность шагов, предвидя результат каждого шага и оценивая их влияние на внешнее окружение планируемой деятельности на взаимоотношения участников этой деятельности	Тест	Выполнение теста на 70-100%	Выполнение менее 70%
	Знает как осуществляется поиск вариантов решения поставленной задачи на основе доступных источников информации; определяет в рамках выбранного алгоритма вопросы, подлежащие дальнейшей разработке	Решение стандартных практических задач	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены
	Владеет навыками анализа проблемной ситуации как системы, выявляя ее	Решение прикладных задач в конкретной предметной области	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены

	составляющие и связи между ними			
УК-2	Знать принципы формирования результатов проекта (или отдельных его этапов) в форме отчета, статьи, выступления на научнопрактической конференции	Тест	Выполнение теста на 70-100%	Выполнение менее 70%
	Уметь организовать и координировать работу участников проекта, обеспечивает работу команды необходимыми ресурсами	Решение стандартных практических задач	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены
	Владеть навыками формирования плана-графика реализации проекта в целом и контролирует его выполнение	Решение прикладных задач в конкретной предметной области	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены
ОПК-10	Знать методологии разработки программного обеспечения, методологии управления проектами разработки программного обеспечения, методы и средства организации проектных данных, лучшие практики управления разработкой программного обеспечения, нормативно-технические документы (стандарты и регламенты), описывающие процессы управления инфраструктурой коллективной среды разработки	Тест	Выполнение теста на 70-100%	Выполнение менее 70%
	Уметь применять методологии разработки программного обеспечения, методологии управления проектами разработки программного обеспечения, методы и средства организации проектных данных, лучшие практики и отражать их в базе знаний, основные принципы и методы управления персоналом, нормативно-технические документы (стандарты и регламенты), описывающие процессы управления инфраструктурой	Решение стандартных практических задач	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены

	коллективной среды разработки			
	Владеть навыками выбора инструментальных средств разработки, определения набора библиотек повторно используемых модулей, выбора средств создания и ведения репозитория, учета задач, сборки и непрерывной интеграции, базы знаний	Решение прикладных задач в конкретной предметной области	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены
ПК-2	Знать сквозные цифровые технологии искусственного интеллекта, включая «Компьютерное зрение», «Обработка естественного языка», «Рекомендательные системы и системы поддержки принятия решений», «Распознавание и синтез речи»; современный опыт применения систем искусственного интеллекта	Тест	Выполнение теста на 70-100%	Выполнение менее 70%
	Уметь проводить сравнительный анализ методов и инструментальных средств для решения задач искусственного интеллекта; разрабатывать и оценивать модели искусственного интеллекта; программировать на языках высокого уровня, ориентированных на применение методов и алгоритмов искусственного интеллекта.	Решение стандартных практических задач	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены
	Владеть навыками выявления требований заказчика к результатам решения задач искусственного интеллекта; определения возможностей применения методов искусственного интеллекта в предметной области решаемой задачи; использования имеющейся методологической и технологической инфраструктуры	Решение прикладных задач в конкретной предметной области	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены

	анализа и обработки данных.			
--	-----------------------------	--	--	--

7.2 Примерный перечень оценочных средств (типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности)

7.2.1 Примерный перечень заданий для подготовки к тестированию

1. Какой из следующих методов можно использовать для нормализации ключевых слов в NLP, процесс преобразования ключевого слова в его базовую форму

- +а) Лемматизация
- б) Soundex
- с) Косинус сходства
- д) N-грамм

2. N-граммы определяются как комбинация N ключевых слов вместе.

Сколько биграмм можно составить из данного предложения: “Analytics Vidhya is a great source to learn data science”

- а)7
- б)8
- +с)9
- д)10

3. Какой из следующих методов можно использовать для вычисления расстояния между двумя векторами слов в NLP?

- а) Лемматизация
- +б) Евклидово расстояние
- +с) Косинус сходства
- д) N-грамм

4. Каковы возможные особенности корпуса текстов в NLP?

- а) Количество слов в документе
- б) Векторное обозначение слова
- с) Часть речевого тега
- д) Базовая грамматика зависимостей
- е) Все вышеперечисленное

4. Вы создали матрицу терминов документа на основе входных данных 20 000 документов для модели машинного обучения. Что из перечисленного можно использовать для уменьшения размерности данных?

- 1 Нормализация ключевых слов.
- 2 Скрытое семантическое индексирование.
- 3 Скрытое распределение Дирихле.

- А) только 1
- Б) 2, 3
- В) 1, 3
- +Г) 1, 2, 3

5. Какой из методов синтаксического анализа текста можно использовать для обнаружения именной фразы, глагольной фразы, обнаружения субъекта и обнаружения объекта в NLP.

- а. Тегирование части речи

- б. Пропустить извлечение граммов и N-граммов
- в. Непрерывный мешок слов
- +г. Анализ зависимостей и анализ групп
- 6. Различие между словами, выраженное с помощью косинусного сходства, будет иметь значения, значительно превышающие 0,5.
 - а. Истинный
 - б. Ложь
- 7. Что из нижеперечисленного относится к методам нормализации ключевых слов в NLP?
 - +а. Стемминг
 - б. Часть речи
 - в. Распознавание именованных объектов
 - +г. Лемматизация
- 8. Что из нижеперечисленного относится к вариантам использования NLP?
 - а. Обнаружение объектов на изображении
 - б. Распознавание лиц
 - в. Речь биометрическая
 - +г. Обобщение текста
- 9. В корпусе из N документов один случайно выбранный документ содержит в общей сложности T терминов, а термин «привет» встречается K раз. Каково правильное значение произведения TF (частота терминов) и IDF (обратная частота документа), если термин «привет» встречается примерно в одной трети всех документов?
 - а. $KT * \text{Log}(3)$
 - б. $T * \text{Log}(3) / K$
 - +в. $K * \text{Log}(3) / T$
 - г. $\text{Log}(3) / KT$
- 10. В NLP алгоритм уменьшает вес часто используемых слов и увеличивает вес слов, которые редко используются в наборе документов.
 - а. Термин Частота (TF)
 - +б. Обратная частота документа (IDF)
 - в. Word2Vec
 - г. Скрытое распределение Дирихле (LDA)

7.2.2 Примерный перечень заданий для решения стандартных задач

задач

1. В NLP процесс удаления таких слов, как «и», «является», «а», «ан», «то» из предложения называется
 - а. Стемминг
 - б. лемматизация
 - +в. Стоп-слово
 - г. Все вышеперечисленное
2. В NLP процесс преобразования предложения или абзаца в токены называется стеммингом.
 - а. Истинный

- +б. Ложь
- 3. В NLP токены преобразуются в числа перед передачей в любую нейронную сеть.
 - +а. Истинный
 - б. Ложь
- 4. определить лишнее
 - а. nltk
 - б. scikit learn
 - в. SpaCy
 - +г. BERT
- 5. TF-IDF помогает вам установить?
 - а. наиболее часто встречающееся слово в документе
 - +б. самое важное слово в документе
- 6. В NLP процесс идентификации людей, организации из заданного абзаца предложения называется
 - а. Стемминг
 - б. Лемматизация
 - в. Удаление стоп-слов
 - +д. Распознавание именованных объектов
- 7. Что из нижеперечисленного не относится к методам предварительной обработки в NLP?
 - а. Стемминг и лемматизация
 - б. Преобразование в нижний регистр
 - в. Удаление знаков препинания
 - г. Удаление стоп-слов
 - +д. Анализ настроений
- 8. При анализе текста преобразование текста в токены, а затем преобразование их в целочисленные векторы или векторы с плавающей запятой можно выполнить с помощью
 - +а. CountVectorizer
 - б. TF-IDF
 - в. Мешок слов
 - г. NER
- 9. В NLP слова, представленные в виде векторов, называются нейронными вложениями слов.
 - +а. Истинный
 - б. Ложь
- 10. В NLP поддерживается контекстное моделирование, с помощью которого одно из следующих вложений слов
 - а. Word2Vec
 - б. GloVe
 - +в. BERT
 - г. Все вышеперечисленное

7.2.3 Примерный перечень заданий для решения прикладных задач

1. В NLP двунаправленный контекст поддерживается каким из следующих вложений
 - а. Word2Vec
 - +б. BERT
 - в. GloVe
 - г. Все вышеперечисленное
2. Какое из следующих вложений Word может быть специально обучено для определенного предмета в NLP
 - а. Word2Vec
 - +б. BERT
 - в. GloVe
 - г. Все вышеперечисленное
3. Вложения Word охватывают несколько измерений данных и представляются в виде векторов.
 - +а. Истинный
 - б. Ложь
4. В NLP векторы встраивания Word помогают установить расстояние между двумя токенами.
 - +а. Истинный
 - б. Ложь
5. Языковые предубеждения вводятся из-за исторических данных, используемых во время обучения встраиванию слов, которое из приведенных ниже не является примером предвзятости.
 - +а. Нью-Дели — в Индию, Пекин — в Китай
 - б. Мужчина для компьютера, женщина для домохозяйки
 - б. Что из следующего будет лучшим выбором для решения таких случаев использования NLP, как семантическое сходство, понимание прочитанного и рассуждения здравого смысла
 - а. GloVe
 - б. BERT
 - +в. Open AI's GPT
 - г. ULMFit
7. Какая из следующих архитектур может быть обучена быстрее и требует меньшего количества обучающих данных
 - а. Языковое моделирование на основе LSTM
 - +б. Трансформаторная архитектура
8. Одно и то же слово может иметь несколько вложений слов с помощью _____?
 - а. GloVe
 - б. Word2Vec
 - +в. ELMo
 - г. nltk
9. Для данного токена его входное представление представляет собой сумму встраивания из токена, сегмента и встраивания позиции.
 - а. ELMo

- б. GPT
- +в. BERT
- г. ULMFit

10. Обучает две независимые языковые модели LSTM слева направо и справа налево и неглубоко объединяет их

- а. GPT
- б. BERT
- в. ULMFit
- +г. ELMo

7.2.4 Примерный перечень вопросов для подготовки к зачету

- 1 Типы задач обработки естественного языка.
- 2 Традиционные методы и методы глубинного обучения.
- 3 Представления для слов естественного языка.
- 4 Векторное представление.
- 5 Целевая функция векторизации слов.
- 6 Построение векторизованной модели словаря.
- 7 Сокращение размерности словаря.
- 8 Методы обучения.
- 9 Задачи классификации.
- 10 Учет структуры предложения.
- 11 Безконтекстная грамматика.
- 12 Проблема неопределенности модификатора.
- 13 Типы зависимостей.
- 14 Статистическая модель языка.
- 15 N-граммы.
- 16 Рекуррентные нейронные сети.
- 17 Оценка качества перевода.
- 18 Отображение последовательностей (seq2seq).
- 19 Свёрточные нейронные сети.
- 20 Применение в системах анализа естественного языка.
- 21 Работа с частями слов.
- 22 Буквенные (character-based) модели.
- 23 Гибридные модели.
- 24 Контекстные представления.
- 25 Проблема неизвестных слов.
- 26 Механизмы внимания, их реализация, применение.
- 27 Самовнимание.
- 28 Трансформеры.
- 29 Перенесение обучения между языковыми моделями.
- 30 Обзор применения технологий работы с естественным языком.

7.2.5 Примерный перечень заданий для подготовки к экзамену

Не предусмотрено учебным планом

7.2.6. Методика выставления оценки при проведении промежуточной аттестации

Зачет проводится по тест-билетам, каждый из которых содержит 10 вопросов и задачу. Каждый правильный ответ на вопрос в тесте оценивается 1 баллом, задача оценивается в 10 баллов (5 баллов верное решение и 5 баллов за верный ответ). Максимальное количество набранных баллов – 20.

1. Оценка «Не зачтено» ставится в случае, если студент набрал менее 11 баллов.

2. Оценка «Зачтено» ставится в случае, если студент набрал от 11 до 20 баллов.)

7.2.7 Паспорт оценочных материалов

№ п/п	Контролируемые разделы (темы) дисциплины	Код контролируемой компетенции	Наименование оценочного средства
1	Обработка естественного языка	УК-1, УК-2, ПК-8, ПК-5	Тест, контрольная работа, защита лабораторных работ, защита реферата, требования к курсовому проекту....
2	Векторное представление.	УК-1, УК-2, ПК-8, ПК-5	Тест, контрольная работа, защита лабораторных работ, защита реферата, требования к курсовому проекту....
3	Методы обучения.	УК-1, УК-2, ПК-8, ПК-5	Тест, контрольная работа, защита лабораторных работ, защита реферата, требования к курсовому проекту....
4	Рекуррентные нейронные сети	УК-1, УК-2, ПК-8, ПК-5	Тест, контрольная работа, защита лабораторных работ, защита реферата, требования к курсовому проекту....
5	Свёрточные нейронные сети	УК-1, УК-2, ПК-8, ПК-5	Тест, контрольная работа, защита лабораторных работ, защита реферата, требования к курсовому проекту....
6	Механизмы внимания	УК-1, УК-2, ПК-8, ПК-5	Тест, контрольная работа, защита лабораторных работ, защита реферата, требования к курсовому проекту....

7.3. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Тестирование осуществляется, либо при помощи компьютерной системы тестирования, либо с использованием выданных тест-заданий на бумажном носителе. Время тестирования 30 мин. Затем осуществляется проверка теста экзаменатором и выставляется оценка согласно методики выставления оценки при проведении промежуточной аттестации.

Решение стандартных задач осуществляется, либо при помощи компьютерной системы тестирования, либо с использованием выданных задач на бумажном носителе. Время решения задач 30 мин. Затем осуществляется проверка решения задач экзаменатором и выставляется оценка, согласно методики выставления оценки при проведении промежуточной аттестации.

Решение прикладных задач осуществляется, либо при помощи компьютерной системы тестирования, либо с использованием выданных задач на бумажном носителе. Время решения задач 30 мин. Затем осуществляется проверка решения задач экзаменатором и выставляется оценка, согласно методики выставления оценки при проведении промежуточной аттестации.

8 УЧЕБНО МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ)

8.1 Перечень учебной литературы, необходимой для освоения дисциплины

1 Python и анализ данных [Электронный ресурс] / Маккинли Уэс, А. Слинкина ; Уэс Маккинли; пер. А. Слинкина. - Python и анализ данных ; 2024-10-28. - Саратов : Профобразование, 2019. - 482 с.

2 Анализ данных [Электронный ресурс] : Учебно-методическое пособие / Г. В. Шнарева, Ж. Г. Пономарева ; Г. В. Шнарева, Ж. Г. Пономарева. - Анализ данных ; 2024-12-06. - Симферополь : Университет экономики и управления, 2019. - 129 с. –

8.2 Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень лицензионного программного обеспечения, ресурсов информационно-телекоммуникационной сети «Интернет», современных профессиональных баз данных и информационных справочных систем:

Лицензионное ПО:

- Microsoft Word
- Intellij PyCharm Professional

Свободное программное обеспечение:

- LibreOffice
- Jupyter Notebook
- Python

Отечественное ПО:

- СУБД Линтер

Ресурсы информационно-телекоммуникационной сети «Интернет»:

- <http://www.edu.ru/>

- Образовательный портал ВГТУ

Информационные справочные системы:

- <http://window.edu.ru>
- <https://wiki.cchgeu.ru/>

9 МАТЕРИАЛЬНО-ТЕХНИЧЕСКАЯ БАЗА, НЕОБХОДИМАЯ ДЛЯ ОСУЩЕСТВЛЕНИЯ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА

Специализированная лекционная аудитория, оснащенная оборудованием для лекционных демонстраций и проекционной аппаратурой.

Учебные лаборатории (г. Воронеж, ул. Плехановская, д. 11):

- 202/2.

- 215/2.

Дисплейный класс, оснащенный компьютерными программами для проведения лабораторного практикума.

Кабинеты, оборудованные проекторами и интерактивными досками.

10. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)

По дисциплине «Технологии работы с естественным языком Nature Language Processing» читаются лекции, проводятся лабораторные работы.

Основой изучения дисциплины являются лекции, на которых излагаются наиболее существенные и трудные вопросы, а также вопросы, не нашедшие отражения в учебной литературе.

Лабораторные работы выполняются на лабораторном оборудовании в соответствии с методиками, приведенными в указаниях к выполнению работ.

Вид учебных занятий	Деятельность студента
Лекция	Написание конспекта лекций: кратко, схематично, последовательно фиксировать основные положения, выводы, формулировки, обобщения; пометать важные мысли, выделять ключевые слова, термины. Проверка терминов, понятий с помощью энциклопедий, словарей, справочников с выписыванием толкований в тетрадь. Обозначение вопросов, терминов, материала, которые вызывают трудности, поиск ответов в рекомендуемой литературе. Если самостоятельно не удастся разобраться в материале, необходимо сформулировать вопрос и задать преподавателю на лекции или на практическом занятии.
Лабораторная работа	Лабораторные работы позволяют научиться применять теоретические знания, полученные на лекции при решении конкретных задач. Чтобы наиболее рационально и полно использовать все возможности лабораторных для подготовки к ним необходимо: следует разобрать лекцию по соответствующей теме, ознакомиться с соответствующим разделом учебника, проработать дополнительную литературу и источники, решить задачи и выполнить другие письменные задания.
Самостоятельная работа	Самостоятельная работа студентов способствует глубокому усвоения учебного материала и развитию навыков самообразования. Самостоятельная работа предполагает следующие составляющие: <ul style="list-style-type: none">- работа с текстами: учебниками, справочниками, дополнительной литературой, а также проработка конспектов лекций;- выполнение домашних заданий и расчетов;- работа над темами для самостоятельного изучения;- участие в работе студенческих научных конференций, олимпиад;- подготовка к промежуточной аттестации.

Подготовка к промежуточной аттестации	Готовиться к промежуточной аттестации следует систематически, в течение всего семестра. Интенсивная подготовка должна начаться не позднее, чем за месяц-полтора до промежуточной аттестации. Данные перед зачетом три дня эффективнее всего использовать для повторения и систематизации материала.
---------------------------------------	---